Statistical Language

***Statistical Language** helps you to understand a range of statistical concepts and terms with simple explanations.*

Find concept definitions:

Statistical Language Glossary

Explore a concept:

What are Data?
- Data unit
- Data item (variable)
- Observation
- Dataset

Quantitative and Qualitative Data
- Quantitative data
- Qualitative data

What are Variables?
- Variable (data item)
- Numeric
- Continuous
- Discrete
- Categorical
- Ordinal
- Nominal

What is a Population?
- Population

Census and Sample
- Census
- Sample
- Random (probability) sample
- Non-random (non-probability) sample

Data Sources
- Direct/Primary data
- Survey
- Indirect/Secondary data
- Administrative data

Describing Frequencies
- Absolute frequency
- Relative frequency
- Ratio
- Rate
- Proportion

Frequency Distribution
- Frequency distribution
- Histogram
- Bar chart

Measures of Shape
- Measures of shape
- Normal distribution
- Skewness

Measures of Central Tendency
- Mode
- Median
- Mean
- Outlier

Measures of Spread
- Range
- Quartiles
- Interquartile range
- Variance
- Standard deviation

Types of Error
- Sampling error
- Non-sampling error

Measures of Error
- Standard error (SE)
- Relative error (RE)
- Confidence interval

What are Statistics?
- Descriptive (summary) statistics
- Inferential statistics

What are Standards?
- Statistical standard
- Classification

What is Metadata?
- Metadata

Data Visualisation
- Static
- Dynamic
- Interactive

Time Series Data
- Original time series
- Seasonally adjusted time series
- Trend series

Estimate and Projection
- Estimate
- Projection

Correlation and Causation
- Correlation
- Causation (Causality)

Confidentiality
- Confidentiality

This page last updated 29 May 2013

What are Data?

This animation explains the concept of data. If you are unable to access the video a Transcript (.doc 29kb) *has been provided. The animation requires* Adobe Flash Player *to run. The animation contains no audio.*

What are data?

Data are measurements or observations that are collected as a source of information. There are a variety of different types of data, and different ways to represent data.

The number of people in Australia, the countries where people were born, number of calls received by the emergency services each day, the value of sales of a particular product, or the number of times Australia has won a cricket match, are all examples of data.

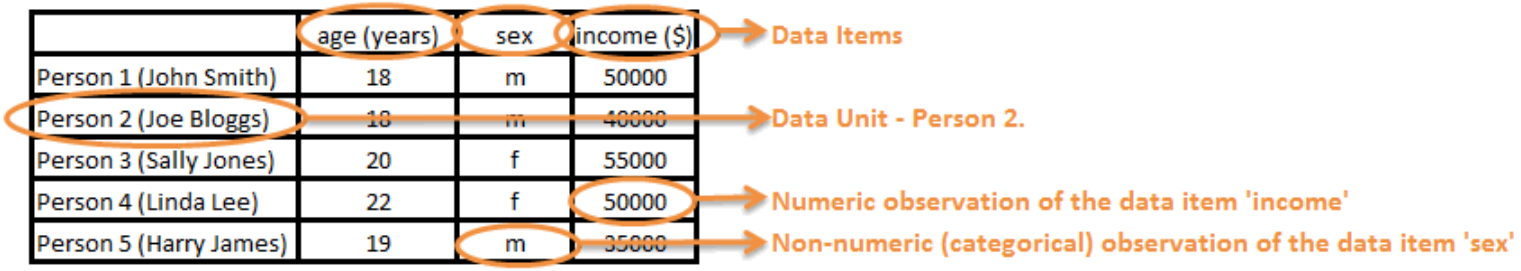
A data unit is one entity (such as a person or business) in the population being studied, about which data are collected. A data unit is also referred to as a unit record or record.

A data item is a characteristic (or attribute) of a data unit which is measured or counted, such as height, country of birth, or income. A data item is also referred to as a **variable** because the characteristic may vary between data units, and may vary over time.

An observation is an occurrence of a specific data item that is recorded about a data unit. It may also be referred to as **datum**, which is the singular form of data. An observation may be numeric or non-numeric (categorical). For example, 173 is a numeric observation of the data item 'height (cm)', whereas 'Australia' is a non-numeric (categorical) observation of the data item 'country of birth'.

A dataset is a complete collection of all observations.

The following table is an example of a dataset:



Return to Statistical Language Homepage

This page last updated 3 July 2013

Statistical Language - Quantitative and Qualitative Data

Quantitative and Qualitative Data

This animation explains the concept of quantitative and qualitative data. If you are unable to access the video a Transcript (.doc 55kb) *has been provided. The animation requires* Adobe Flash Player *to run. The animation contains no audio.*

What are quantitative and qualitative data?

Quantitative data are measures of values or counts and are expressed as numbers.

Quantitative data are data about **numeric variables** (e.g. how many; how much; or how often).

Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.

Qualitative data are data about **categorical variables** (e.g. what type).

Quantitative = Quantity	Qualitative = Quality

Data collected about a numeric variable will always be quantitative and data collected about a categorical variable will always be qualitative. Therefore, you can identify the type of data, prior to collection, based on whether the variable is numeric or categorical.

Why are quantitative and qualitative data important?

Quantitative and qualitative data provide different outcomes, and are often used together to get a full picture of a population. For example, if data are collected on annual income (quantitative), occupation data (qualitative) could also be gathered to get more detail on the average annual income for each type of occupation.

Quantitative and qualitative data can be gathered from the same data unit depending on whether the variable of interest is numerical or categorical. For example:

Data unit	Numeric variable	= Quantitative data	Categorical variable	= Qualitative data
A person	"*How many* children do you have?"	**4** children	"In *which country* were your children born?"	**Australia**
	"*How much* do you earn?"	$**60,000** p.a.	"*What* is your occupation?"	**Photographer**
	"How many hours do you work?"	**38** hours per week	"Do you work *full-time or part-time*?"	**Full-time**
A house	"*How many* square metres is the house?"	**200** square metres	"In *which city or town* is the house located?"	**Brisbane**
A business	"*How many* workers are currently employed?"	**264** employees	"*What* is the *industry* of the business?"	**Retail**
A farm	"*How many* milk cows are located on the farm?	**36** cows	"*What* is the main *activity* of the farm?"	**Dairy**

How can you use quantitative and qualitative data?

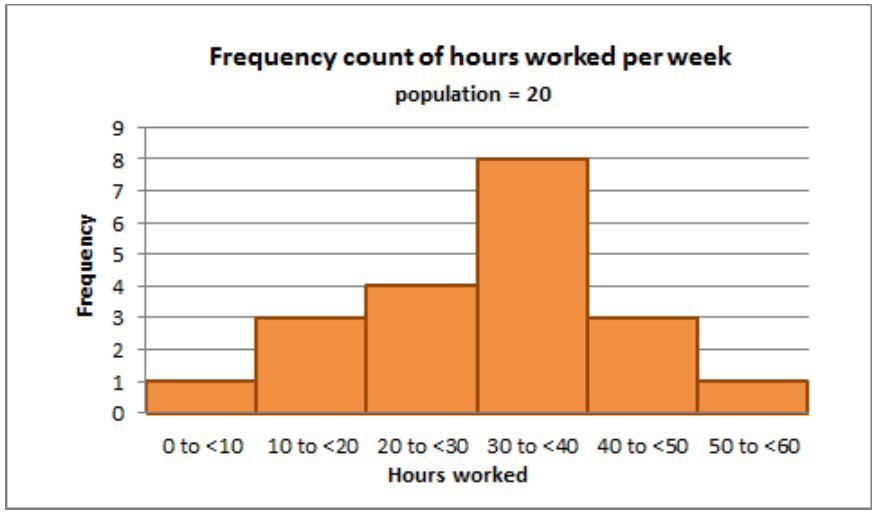
It is important to identify whether the data are quantitative or qualitative as this affects the statistics that can be produced.

Frequency counts:

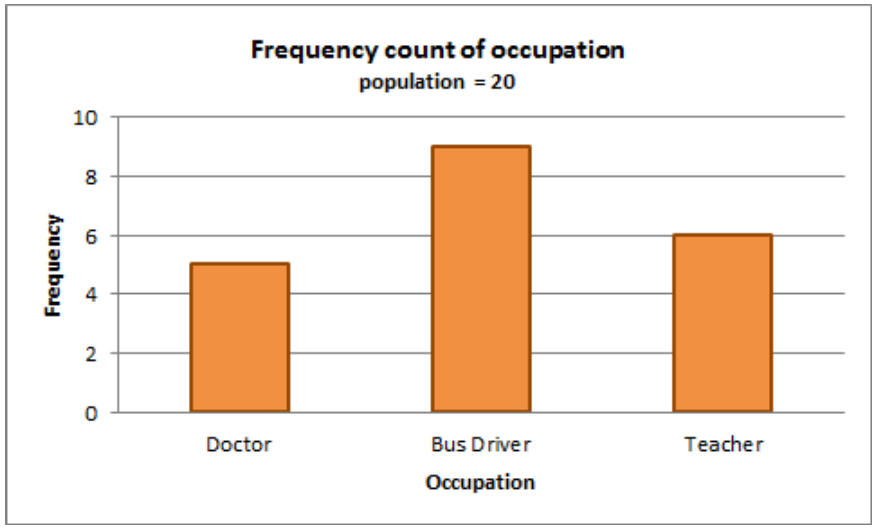
The number of times an observation occurs (frequency) for a data item (variable) can be shown for both quantitative and qualitative data.

The graphs below arrange the quantitative and qualitative data to show the frequency distribution of the data.

Quantitative Data



Qualitative Data



As absolute frequencies can be calculated on quantitative and qualitative data, relative frequencies can also be produced, such as percentages, proportions, rates and ratios. For example, the graphs above show 4 people (20%) worked less than 30 hours per week, and 6 people (30%) are teachers.

Descriptive (summary) statistics:

Statistics that describe or summarise can be produced for quantitative data and to a lesser extent for qualitative data.

As quantitative data are always numeric they can be ordered, added together, and the frequency of an observation can be counted. Therefore, all descriptive statistics can be calculated using quantitative data.

As qualitative data represent individual (mutually exclusive) categories, the descriptive statistics that can be calculated are limited, as many of these techniques require numeric values which can be logically ordered from lowest to highest and which express a count.

Mode can be calculated, as it it the most frequency observed value. Median, measures of shape, measures of spread such as the range and interquartile range require an ordered data set with a logical low-end value and high-end value. Variance and standard deviation require the mean to be calculated, which is not appropriate for categorical variables as they have no numerical value.

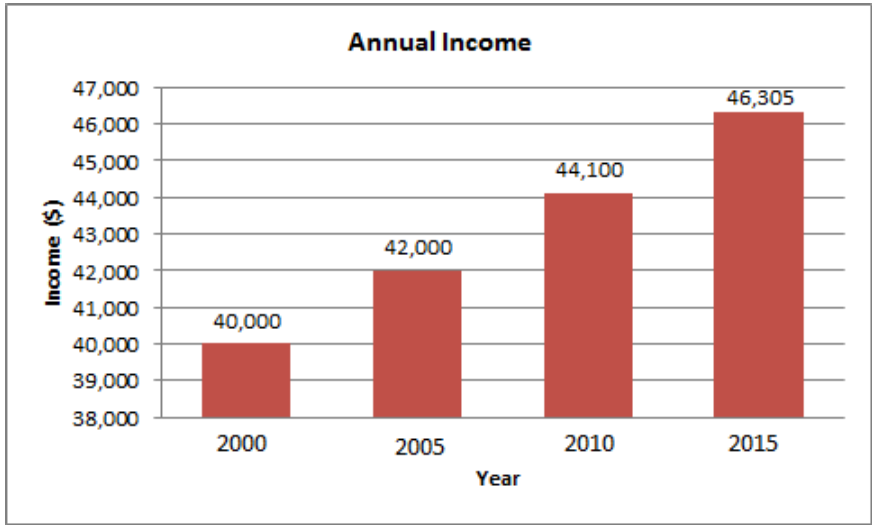
Inferential statistics:

By making inferences about quantitative data from a sample, estimates or projections for the total population can be produced.

Quantitative data can be used to inform broader understandings of a population, or to consider how that population may change or progress into the future.

For example, a simple income projection for an employee in 2015 may be inferred from the rate of change for data collected in 2000, 2005, and 2010.

As shown in the graph below, data collected over time indicates a 5% increase every five years. Therefore, if the rate of increase continues to follow the same pattern, it can be projected that the annual income for that employee in 2015 will be $46,305; which is the 2010 wage of $44,100 increased by an additional 5%.



Qualitative data are not compatible with inferential statistics as all techniques are based on numeric values.

Return to Statistical Language Homepage

This page last updated 4 July 2013

Statistical Language - What are Variables?

What are Variables?

This animation explains the concept of variables. If you are unable to access the video a Transcript (.doc 30kb)*has been provided. The animation requires* Adobe Flash Player *to run. The animation contains no audio.*

What is a variable?

A variable is any characteristics, number, or quantity that can be measured or counted. A variable may also be called a **data item**. Age, sex, business income and expenses, country of birth, capital expenditure, class grades, eye colour and vehicle type are examples of variables. It is called a variable because the value may vary between data units in a population, and may change in value over time.

For example; 'income' is a variable that can vary between data units in a population (i.e. the people or businesses being studied may not have the same incomes) and can also vary over time for each data unit (i.e. income can go up or down).

What are the types of variables?

There are different ways variables can be described according to the ways they can be studied, measured, and presented.

Numeric variables have values that describe a measurable quantity as a number, like 'how many' or 'how much'. Therefore numeric variables are quantitative variables.

Numeric variables may be further described as either continuous or discrete:

- **A continuous variable is a numeric variable. Observations can take any value between a certain set of real numbers.** The value given to an observation for a continuous variable can include values as small as the instrument of measurement allows. Examples of continuous variables include height, time, age, and temperature.

- **A discrete variable is a numeric variable. Observations can take a value based on a count from a set of distinct whole values.** A discrete variable cannot take the value of a fraction between one value and the next closest value. Examples of discrete variables include the number of registered cars, number of business locations, and number of children in a family, all of of which measured as whole units (i.e. 1, 2, 3 cars).

The data collected for a numeric variable are quantitative data.

Categorical variables have values that describe a 'quality' or 'characteristic' of a data unit, like 'what type' or 'which category'. Categorical variables fall into mutually exclusive (in one category or in another) and exhaustive (include all possible options) categories**.** Therefore, categorical variables are qualitative variables and tend to be represented by a non-numeric value.

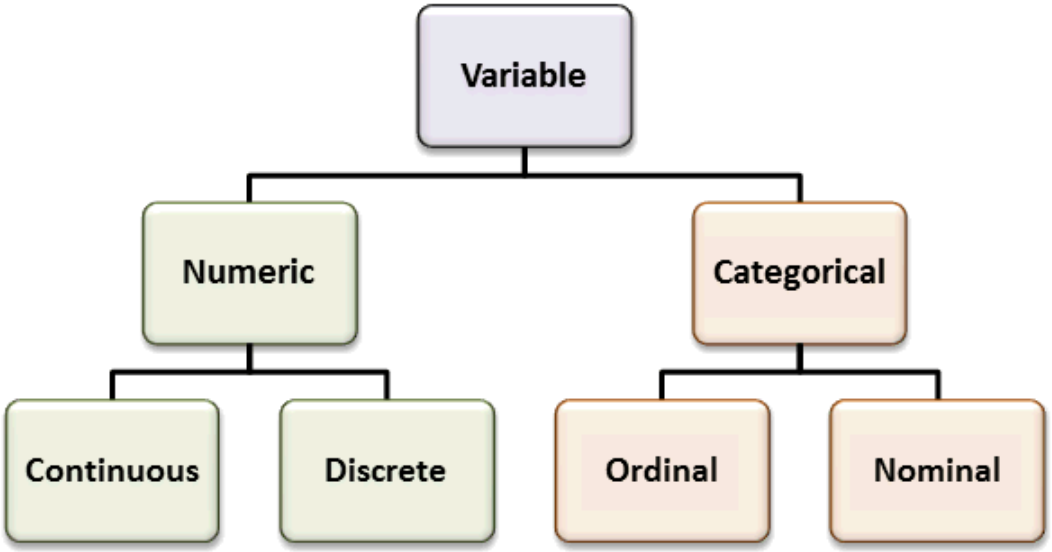
Categorical variables may be further described as ordinal or nominal:

- **An ordinal variable is a categorical variable. Observations can take a value that can be logically ordered or ranked.** The categories associated with ordinal variables can be ranked higher or lower than another, but do not necessarily establish a numeric difference between each category. Examples of ordinal categorical variables include academic grades (i.e. A, B, C), clothing size (i.e. small, medium, large, extra large) and attitudes (i.e. strongly agree, agree, disagree, strongly disagree).

- **A nominal variable is a categorical variable. Observations can take a value that is not able to be organised in a logical sequence.** Examples of nominal categorical variables include sex, business type, eye colour, religion and brand.

The data collected for a categorical variable are qualitative data.

Types of variables flowchart:



Return to Statistical Language Homepage

This page last updated 22 October 2013

Statistical Language - What is a Population?

Statistical Language

What is a Population?

This animation explains the concept of population. If you are unable to access the video a Transcript (.doc 26kb) *has been provided. The animation requires* Adobe Flash Player *to run. The animation contains no audio.*

What is a population?

A population is any complete group with at least one characteristic in common. Populations are not just people. Populations may consist of, but are not limited to, people, animals,

businesses, buildings, motor vehicles, farms, objects or events.

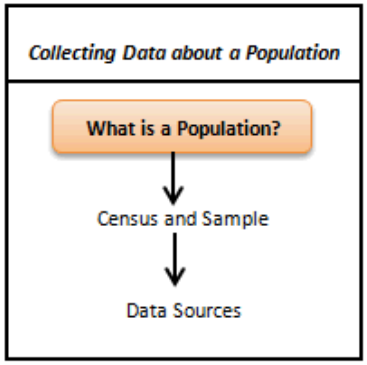
Why do you need to know who or what are in a population?

When looking at data, it is important to clearly identify the population being studied or referred to, so that you can understand *who* or *what* are included in the data. For example, if you were looking at some Australian farming data, you would need to understand whether the population the data refers to is all farms in Australia, just farms that grow crops, those that only have livestock, or some other type of farm.

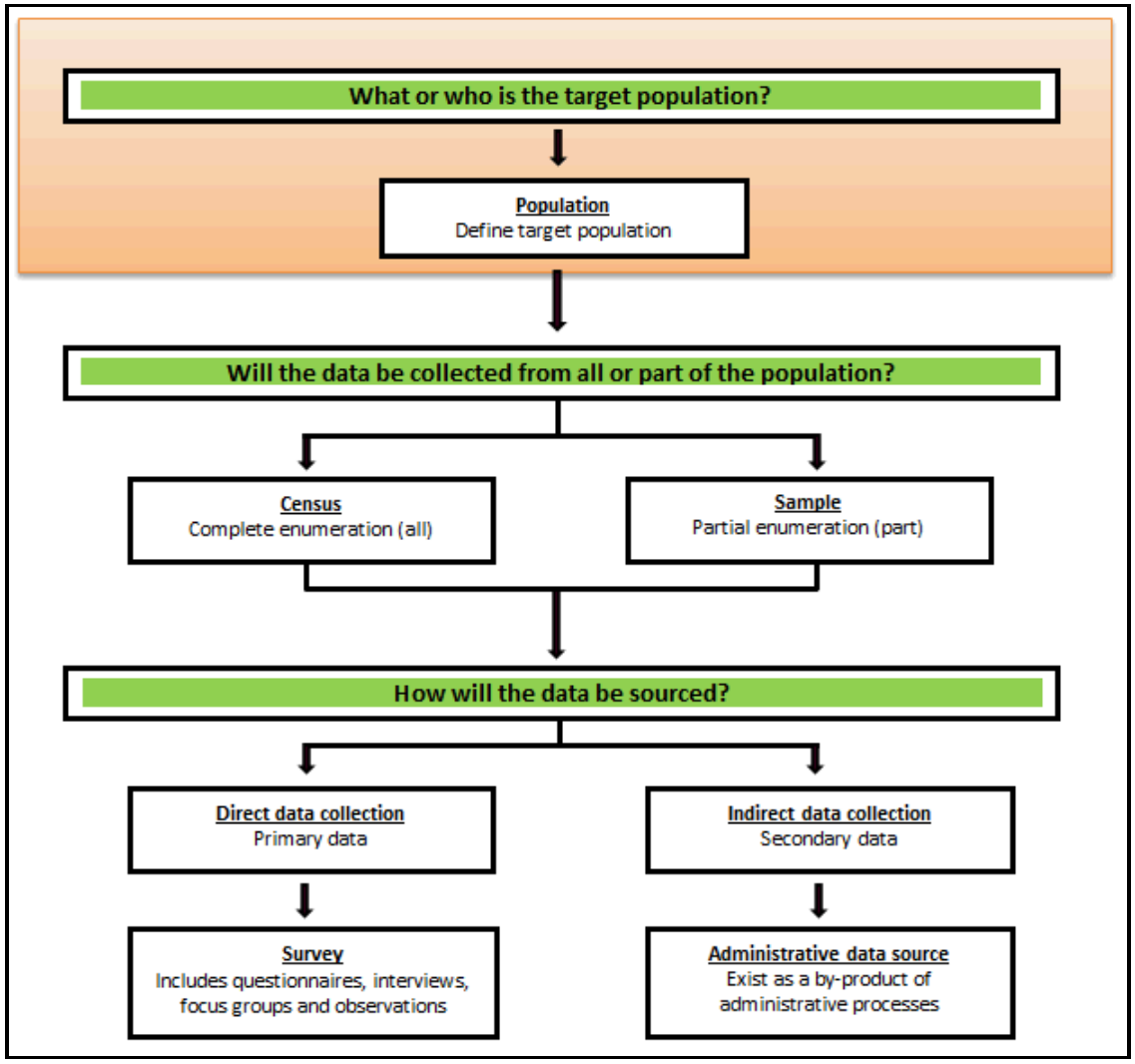
When is a population identified?

The population needs to be clearly identified at the beginning of a study. The study should be based on a clear understanding of who or what is of interest, as well as the type of information required from that population.

Collecting data about a population:



Collecting Data about a Population Flowchart: What is a Population?



Recommended: Read Census and Sample next

Further information:

ABS:
Literacy Stats - Understanding different population counts
3107.0.55.006 - Information Paper: Population Concepts

External links:
Setting up a Survey

Return to Statistical Language Homepage

This page last updated 18 June 2013

Statistical Language - Census and Sample

Census and Sample

Recommended: First read What is a Population?

This animation explains the concept of census and sample. If you are unable to access the video a Transcript (.doc 27kb) *has been provided. The animation requires* Adobe Flash Player *to run. There is no audio in this animation.*

How do we study a population?

A population may be studied using one of two approaches: taking a census, or selecting a sample.

It is important to note that whether a census or a sample is used, both provide information that can be used to draw conclusions about the whole population.

What is a census (complete enumeration)?

A census is a study of every unit, everyone or everything, in a population. It is known as a **complete enumeration**, which means a complete count.

What is a sample (partial enumeration)?

A sample is a subset of units in a population, selected to represent all units in a population of interest. It is a **partial enumeration** because it is a count from part of the population.

Information from the sampled units is used to estimate the characteristics for the entire population of interest.

When to use a census or a sample?

Once a population has been identified a decision needs to be made about whether taking a census or selecting a sample will be the more suitable option. There are advantages and disadvantages to using a census or sample to study a population:

Pros of a CENSUS	Cons of a CENSUS
• provides a true measure of the population (no sampling error) • benchmark data may be obtained for future studies • detailed information about small sub-groups within the population is more likely to be available	• may be difficult to enumerate all units of the population within the available time • higher costs, both in staff and monetary terms, than for a sample • generally takes longer to collect, process, and release data than from a sample
Pros of a SAMPLE	**Cons of a SAMPLE**
• costs would generally be lower than for a census • results may be available in less	• data may not be representative of the total population, particularly where the sample size is small • often not suitable for producing

time • if good sampling techniques are used, the results can be very representative of the actual population	benchmark data • as data are collected from a subset of units and inferences made about the whole population, the data are subject to 'sampling' error • decreased number of units will reduce the detailed information available about sub-groups within a population

How are samples selected?

A sample must be robust in its design and large enough to provide a reliable representation of the whole population. Aspects to be considered when designing a sample include the level of accuracy required, cost, and the timing. Sampling can be random or non-random.

In a random (or probability) sample each unit in the population has a chance of being selected, and this probability can be accurately determined.

Probability or random sampling includes, but is not limited to, simple random sampling, systematic sampling, and stratified sampling. Random sampling makes it possible to produce population estimates from the data obtained from the units included in the sample.

Simple random sample: All members of the sample are chosen at random and have the same chance of being in the sample. A lottery draw is a good example of simple random sampling where the numbers are randomly generated from a defined range of numbers (i.e. 1 through to 45) with each number having an equal chance of being selected.

Systematic random sample: The first member of the sample is chosen at random then the other members of the sample are taken at intervals (i.e. every 4th unit).

Stratified random sample: Relevant subgroups from within the population are identified and random samples are selected from within each strata.

In a non-random (or non-probability) sample some units of the population have no chance of selection, the selection is non-random, or the probability of their selection can not be determined.

In this method the sampling error cannot be estimated, making it difficult to infer population estimates from the sample. Non-random sampling includes convenience sampling, purposive sampling, quota sampling, and volunteer sampling

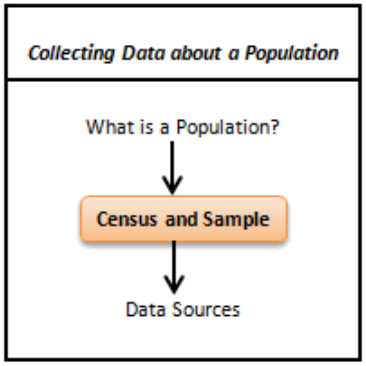
Convenience sampling: Units are chosen based on their ease of access;

Purposive sampling: The sample is chosen based on what the researcher thinks is appropriate for the study;

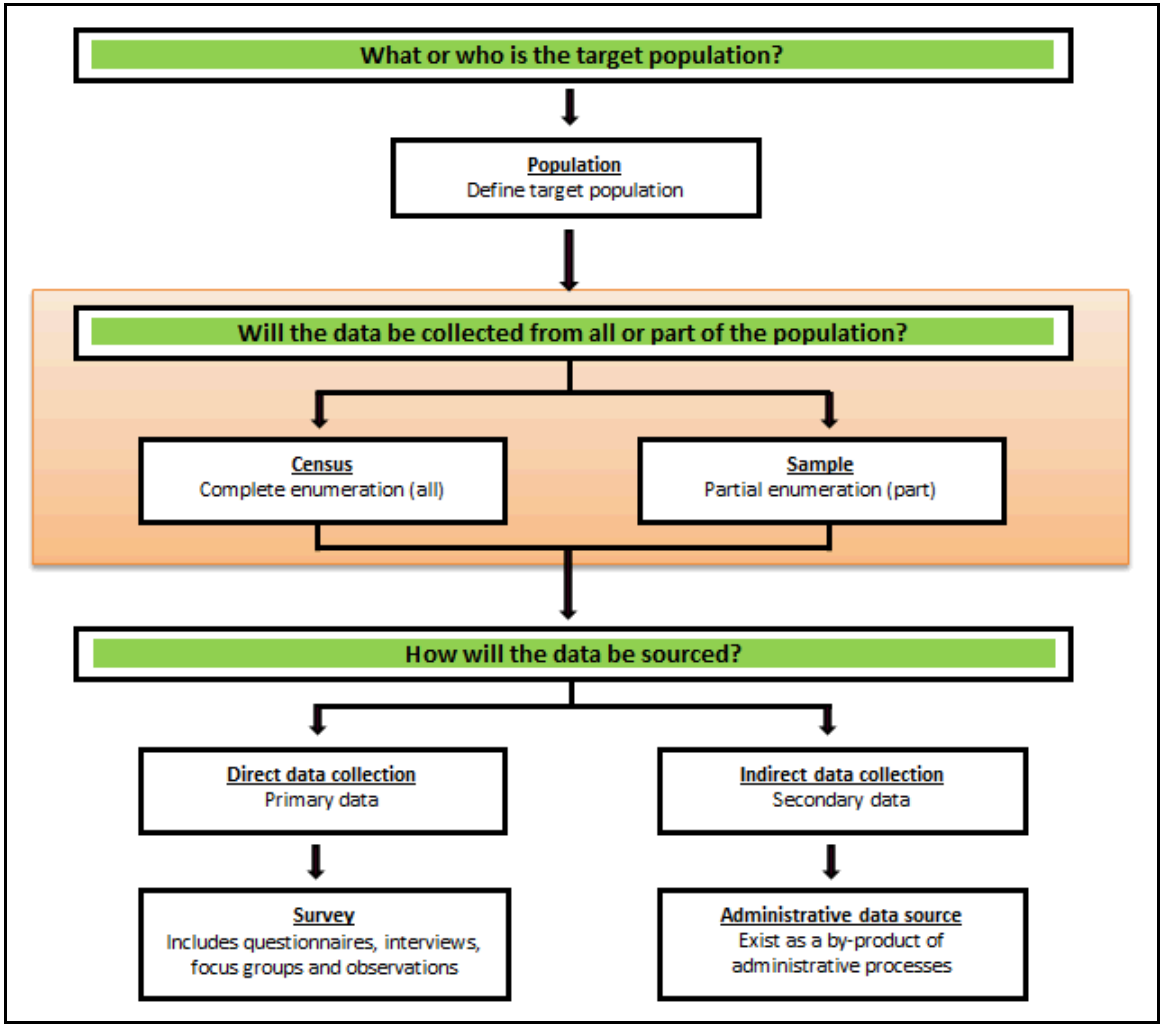
Quota sampling: The researcher can select units as they choose, as long as they reach a defined quota; and

Volunteer sampling: participants volunteer to be a part of the survey (a common method used for internet based opinion surveys where there is no control over how many or who votes).

Collecting data about a population flowchart:



Collecting Data about a Population Flowchart: Census and Sample



Recommended: Read Data Sources next

Further information:

ABS:
1299.0 - An Introduction to Sample Surveys: A User's Guide

External links:

Return to Statistical Language Homepage

This page last updated 3 July 2013

Statistical Language - Data Sources

Data Sources

Recommended: First read Census and Sample

How will the data be sourced?

Data can be sourced directly or indirectly.

Direct methods of data collection involve collecting new data for a specific study. This type of data is known as **primary data**.

Indirect methods of data collection involve sourcing and accessing existing data that were not originally collected for the purpose of the study. This type of data is known as **secondary data**.

How are direct data sourced?

A survey involves collecting information from every unit in the population (a census), or from a subset of units (a sample) from the population.

A respondent provides data about oneself as a unit, or as a representative of another unit in a population.

Methods of direct data collection include:

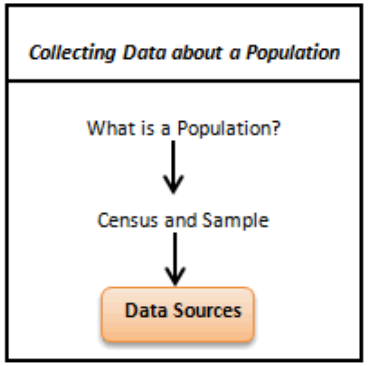
- Surveys administered through the use of an interviewer
- Surveys which are self-enumerated (the information written or entered directly by the respondent)
- In depth interviews or focus groups to provide the opportunity for discussion and elaboration for collecting more detailed information about a particular issue or issues
- Observational studies in which data are gathered through the direct observation of the population or sample
- Experiments and clinical trials that involve controlled studies where researchers collect data from subset groups taken from the population of interest.

How are indirect data sourced?

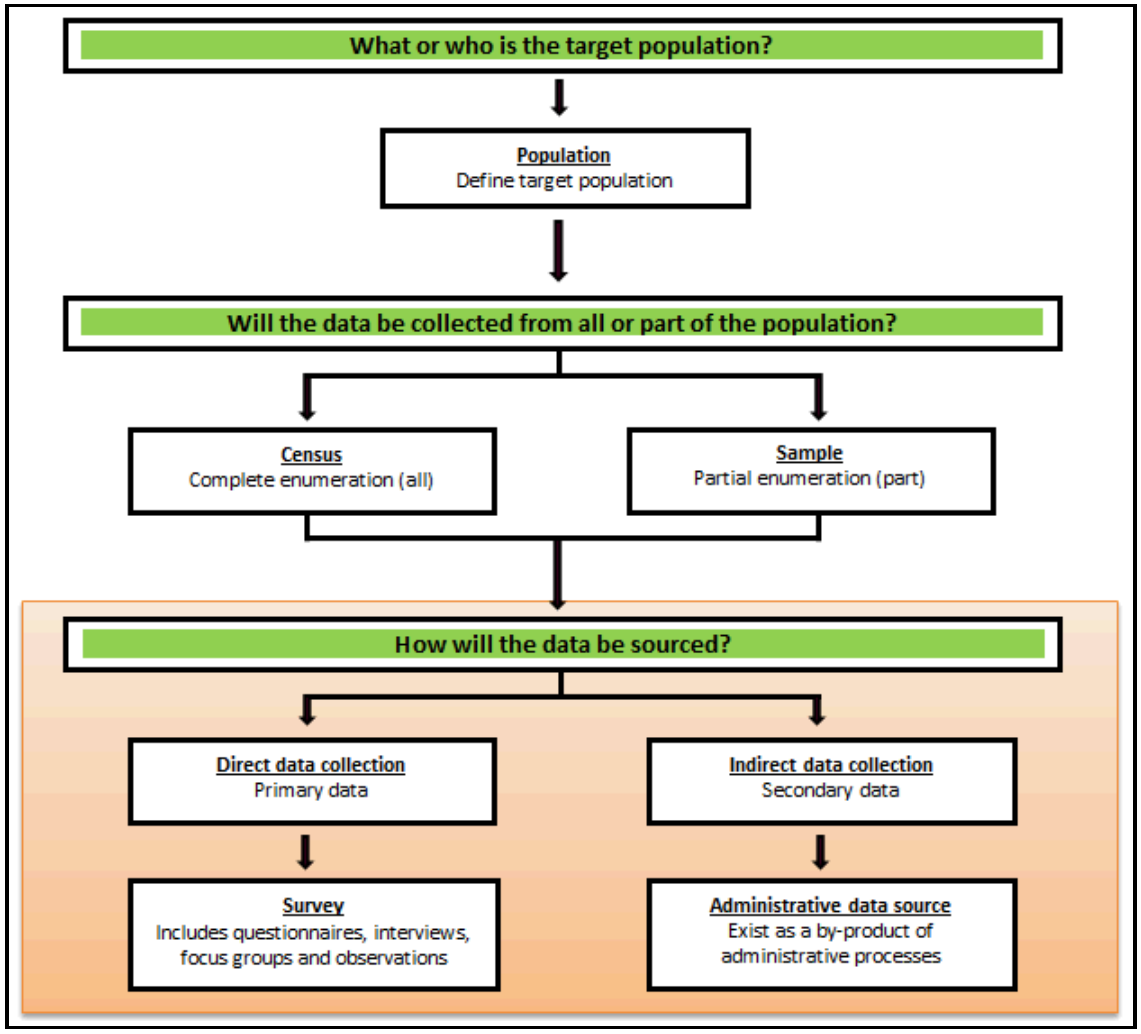
Administrative data are collected as part of the day to day processes and record keeping of organisations. Administrative data, such as historical data or public records, include: school enrolments; hospital admissions; records of births, deaths, and marriages. The data are not collected initially for statistical purposes but can be organised to produce statistics.

Administrative data can be useful because this data are usually recorded about every unit of the population of interest and are continuously collected, allowing comparisons to be made over time. Where administrative data are available, this may eliminate the need to conduct a survey provided the data are fit for the statistical purpose.

Collecting data about a population flowchart:



Collecting Data about a Population Flowchart: Data Sources



Further information:

ABS:
1299.0 - An Introduction to Sample Surveys: A User's Guide
Common Issues with Acquiring Administrative Data

External links:
Surveys and administrative collections
Administrative data as a strategic resource

Return to Statistical Language Homepage

This page last updated 3 July 2013

Statistical Language - Describing Frequencies

Describing Frequencies

This animation explains the concept of frequencies. If you are unable to access the video a Transcript (.doc 28kb) *has been provided. The animation requires* Adobe Flash Player *to run. The animation contains no audio.*

What is a frequency?

The frequency is the number of times a particular value for a variable (data item) has been observed to occur.

How can we measure frequency?

The frequency of a value can be expressed in different ways, depending on the purpose required.

The absolute frequency describes the number of times a particular value for a variable (data item) has been observed to occur.

The simplest way to express a frequency is in absolute terms.

A relative frequency describes the number of times a particular value for a variable (data item) has been observed to occur in relation to the total number of values for that variable.

The relative frequency is calculated by dividing the absolute frequency by the total number of values for the variable.

How are relative frequencies expressed?

Ratios, rates, proportions and percentages are different ways of expressing relative frequencies.

A ratio compares the frequency of one value for a variable with another value for the variable.

The first value identified in a ratio must be to the left of the colon (:) and the second value must be to the right of the colon (1st value : 2nd value).

For example, in a total of 20 coin tosses where there are 12 heads and 8 tails, the ratio of heads to tails is 12:8. Alternatively, the ratio of tails to heads is 8:12.

A rate is a measurement of one value for a variable in relation to another measured quantity.

For example, in a total of 20 coin tosses where there are 12 heads and 8 tails, the rate is 12 heads per 20 coin tosses. Alternatively, the rate is 8 tails per 20 coin tosses.

A proportion describes the share of one value for a variable in relation to a whole.

It is calculated by dividing the number of times a particular value for a variable has been observed, by the total number of values in the population.

For example, in a total of 20 coin tosses where there are 12 heads and 8 tails, the proportion of heads is 0.6 (12 divided by 20). Alternatively, the proportion of tails is 0.4 (8 divided by 20).

A percentage expresses a value for a variable in relation to a whole population as a fraction of one hundred.

The percentage total of an entire dataset should always add up to 100, as 100% represents the total, it is equal to the ‘whole’. A percentage is calculated by dividing the number of times a particular value for a variable has been observed, by the total number of observations in the population, then multiplying this number by 100.

For example, in a total of 20 coin tosses where there are 12 heads and 8 tails, the percentage of heads is 60% (12 divided by 20, multiplied by 100). Alternatively, the percentage of tails is 40% (8 divided by 20, multiplied by 100).

Return to Statistical Language Homepage

This page last updated 3 July 2013

Statistical Language - Frequency Distribution

Frequency Distribution

What is a frequency distribution?

Frequency distributions are visual displays that organise and present frequency counts so that the information can be interpreted more easily.

Frequency distributions can show absolute frequencies or relative frequencies, such as proportions or percentages.

How do we show a frequency distribution?

A frequency distribution of data can be shown in a table or graph. Some common methods of showing frequency distributions include frequency tables, histograms or bar charts.

Frequency Tables

A frequency table is a simple way to display the number of occurrences of a particular value or characteristic.

For example, if we have collected data about height from a sample of 50 children, we could present our findings as:

Height of Children

Height (cm) of children	Absolute frequency	Relative frequency
120 – less than 130	9	18%
130 – less than 140	10	20%
140 – less than 150	13	26%
150 – less than 160	11	22%
160 – less than 170	7	14%

Total	**50**	**100%**

From this frequency table we can quickly identify information such as 7 children (14% of all children) are in the 160 to less than 170 cm height range, and that there are more children with heights in the 140 to less than 150 cm range (26% of all children) than any other height range.

Data can also be presented in graphical form.

Frequency Graphs

Histograms and bar charts are both visual displays of frequencies using columns plotted on a graph. The Y-axis (vertical axis) generally represents the frequency count, while the X-axis (horizontal axis) generally represents the variable being measured.

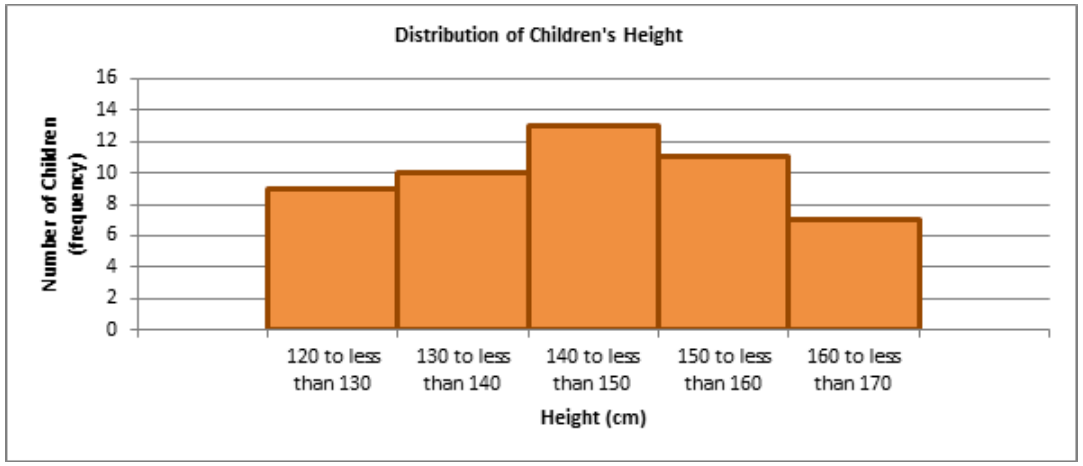
A histogram is a type of graph in which each column represents a numeric variable, in particular that which is continuous and/or grouped.

A histogram shows the distribution of all observations in a quantitative dataset. It is useful for describing the shape, centre and spread to better understand the distribution of the dataset.

Features of a histogram:

- The height of the column shows the frequency for a specific range of values.
- Columns are usually of equal width, however a histogram may show data using unequal ranges (intervals) and therefore have columns of unequal width.
- The values represented by each column must be mutually exclusive and exhaustive. Therefore, there are no spaces between columns and each observation can only ever belong in one column.
- It is important that there is no ambiguity in the labelling of the intervals on the x-axis for continuous or grouped data (e.g. 0 to less than 10, 10 to less than 20, 20 to less than 30).

For example:
The histogram below shows the same information as the frequency table.



A bar chart is a type of graph in which each column (plotted either vertically or horizontally) represents a categorical variable or a discrete ungrouped numeric variable.

It is used to *compare* the frequency (count) for a category or characteristic with another category or characteristic.

Features of a bar chart:

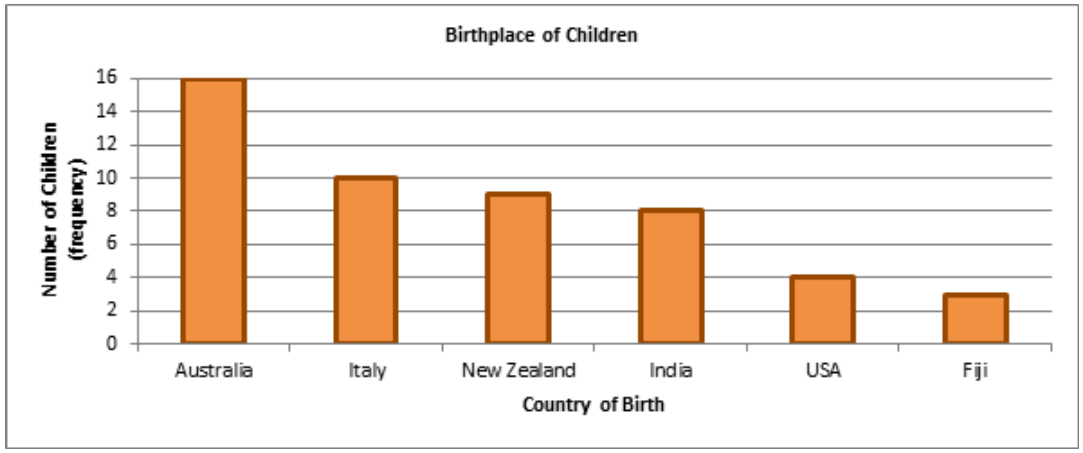
- In a bar chart, the bar height (if vertical) or length (if horizontal) shows the frequency for each category or characteristic.
- The distribution of the dataset is not important because the columns each represent an individual category or characteristic rather than intervals for a continuous measurement. Therefore, gaps are included between each bar and each bar can be arranged in any order without affecting the data.

For example:
If data had been collected for 'country of birth' from a sample of children, a bar chart could be used to plot the data as 'country of birth' is a categorical variable.

Birthplace of Children

Country of Birth	Absolute frequency	Relative frequency
Australia	16	32%
Fiji	3	6%
India	8	16%
Italy	10	20%
New Zealand	9	18%
United States of America	4	8%
Total	**50**	**100%**

The bar chart below shows us that 'Australia' is the most commonly observed country of birth of the 50 children sampled, while 'Fiji' is the least common country of birth.



Return to Statistical Language Homepage

This page last updated 3 July 2013

Statistical Language - Measures of Shape

Measures of Shape

What is a measure of shape?

Measures of shape describe the distribution (or pattern) of the data within a dataset.

The distribution shape of quantitative data can be described as there is a logical order to the values, and the 'low' and 'high' end values on the x-axis of the histogram are able to be identified.

The distribution shape of a qualitative data cannot be described as the data are not numeric.

What are the shapes of a dataset?

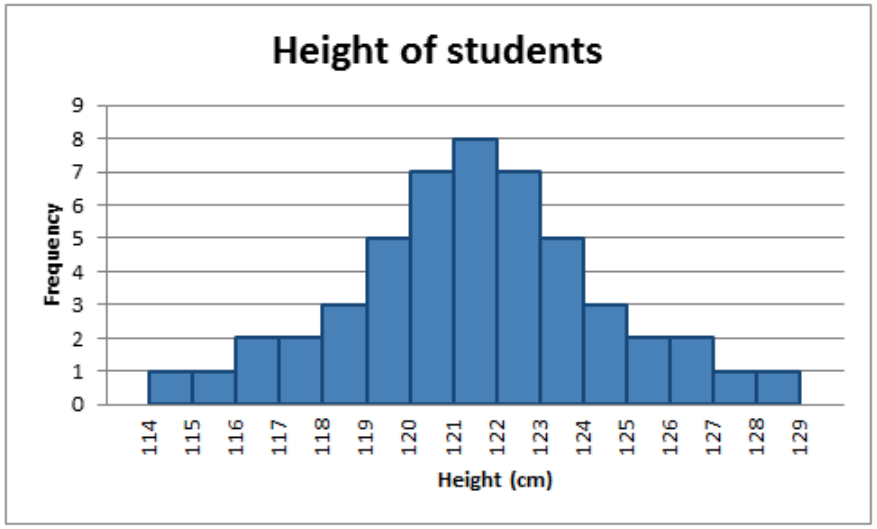
A distribution of data item values may be symmetrical or asymmetrical. Two common examples of symmetry and asymmetry are the 'normal distribution' and the 'skewed distribution'.

In a **symmetrical distribution** the two sides of the distribution are a mirror image of each other.

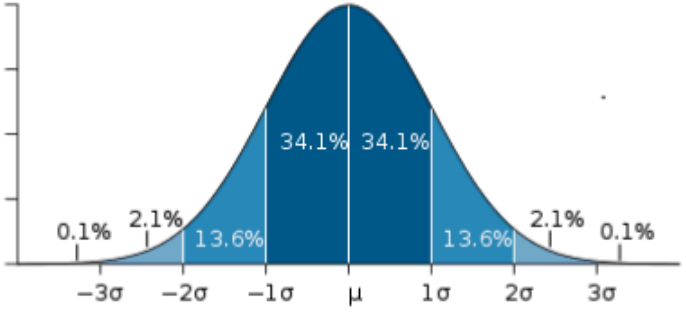
A normal distribution is a true symmetric distribution of observed values.

When a histogram is constructed on values that are normally distributed, the shape of columns form a symmetrical bell shape. This is why this distribution is also known as a 'normal curve' or 'bell curve'.

The following graph is an example of a normal distribution:



If represented as a 'normal curve' (or bell curve) the graph would take the following shape (where μ = mean, and σ = standard deviation):



Key features of the **normal distribution**:

- symmetrical shape
- mode, median and mean are the same and are together in the centre of the curve
- there can only be one mode (i.e. there is only one value which is most frequently observed)
- most of the data are clustered around the centre, while the more extreme values on either side of the centre become less rare as the distance from the centre increases (i.e. About 68% of values lie within one standard deviation (*σ*) away from the mean; about 95% of the values lie within two standard deviations; and about 99.7% are within three standard deviations. This is known as the *empirical rule* or the *3-sigma rule*.)

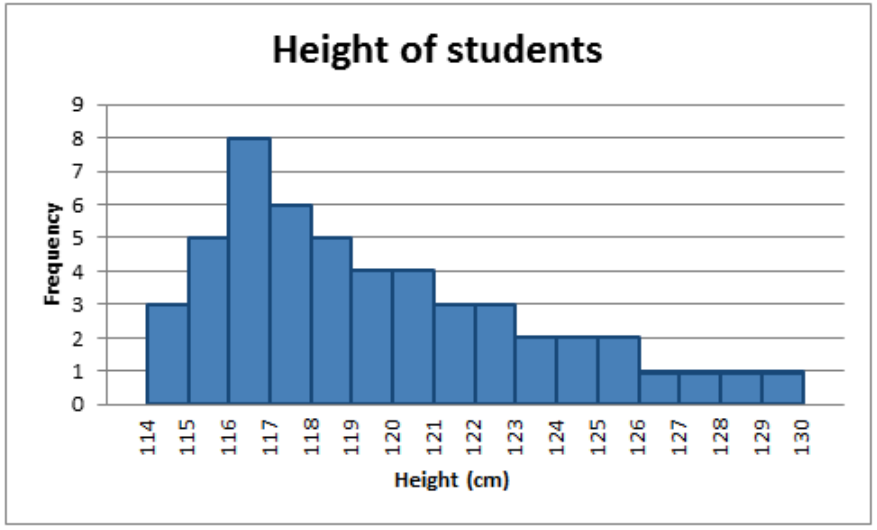
In an **asymmetrical distribution** the two sides will not be mirror images of each other.

Skewness is the tendency for the values to be more frequent around the high or low ends of the x-axis.

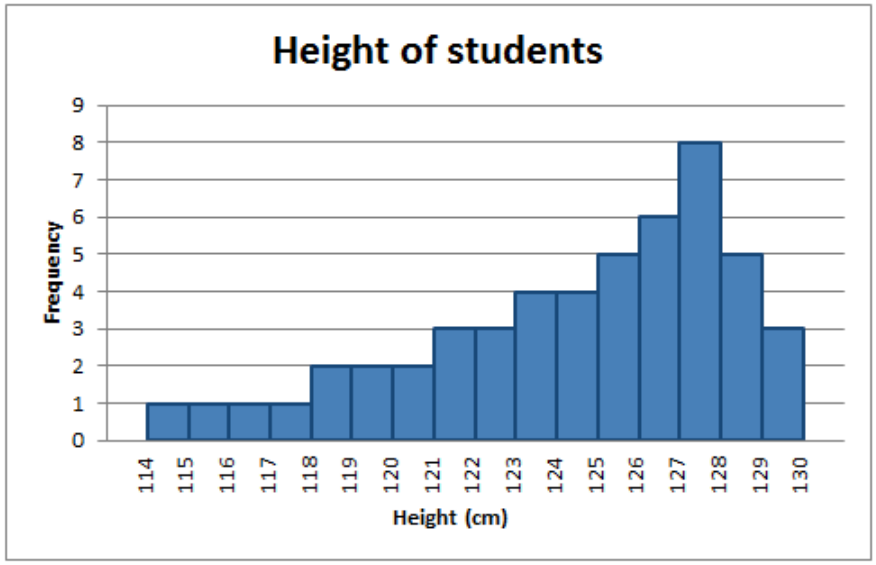
When a histogram is constructed for skewed data it is possible to identify skewness by looking at the shape of the distribution.

For example:

A distribution is said to be **positively skewed** when the tail on the right side of the histogram is longer than the left side. Most of the values tend to cluster toward the left side of the x-axis (i.e. the smaller values) with increasingly fewer values at the right side of the x-axis (i.e. the larger values).



A distribution is said to be **negatively skewed** when the tail on the left side of the histogram is longer than the right side. Most of the values tend to cluster toward the right side of the x-axis (i.e. the larger values), with increasingly less values on the left side of the x-axis (i.e. the smaller values).



Key features of the **skewed distribution**:

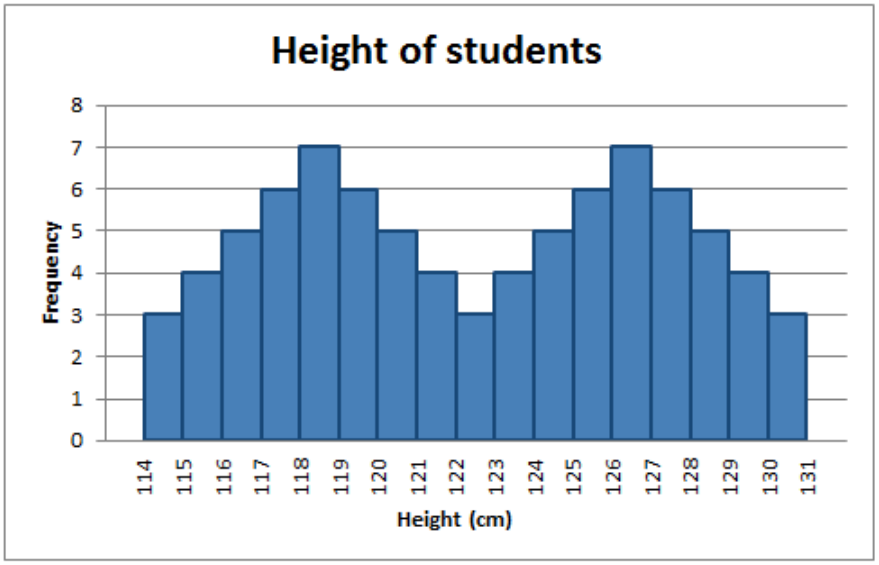
- asymmetrical shape
- mean and median have different values and do not all lie at the centre of the curve
- there can be more than one mode
- the distribution of the data tends towards the high or low end of the dataset

What are the other possible distribution shapes?

Other distributions include uni-modal, bi-modal, or multimodal.

A uni-modal distribution occurs if there is only one 'peak' (a highest point) in the distribution, as seen in the previous histograms. This means there is one mode (a value that occurs more frequently than any other) for the data item (variable).

The distribution shape of the data in the histogram below is bi-modal because there are two modes (two values that occur more frequently than any other) for the data item (variable).



Why are measures of shape useful?

The shape of the distribution can assist with identifying other descriptive statistics, such as which measure of central tendency is appropriate to use.

If the data are normally distributed, the mean, median and mode are all equal, and therefore are all appropriate measure of centre central tendency.

If data are skewed, the median may be a more appropriate measure of central tendency.

Further information:

External links:
Basic Survey Design: Analysis
easycalculation.com - Normal Distribution
easycalculation.com - Skewness calculator

Return to Statistical Language Homepage

This page last updated 3 July 2013

Statistical Language - Measures of Central Tendency

Measures of Central Tendency

Recommended: First read Measures of Shape

What are the measures of central tendency?

A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.

There are three main measures of central tendency: the mode, the median and the mean. Each of these measures describes a different indication of the typical or central value in the distribution.

What is the mode?

The mode is the *most commonly occurring value* in a distribution.

Consider this dataset showing the retirement age of 11 people, in whole years:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

This table shows a simple frequency distribution of the retirement age data.

Age	Frequency
54	3
55	1
56	1

57	2
58	2
60	2

The most commonly occurring value is 54, therefore the mode of this distribution is 54 years.

Advantage of the mode:

The mode has an advantage over the median and the mean as it can be found for both numerical and categorical (non-numerical) data.

Limitations of the mode:

The are some limitations to using the mode. In some distributions, the mode may not reflect the centre of the distribution very well. When the distribution of retirement age is ordered from lowest to highest value, it is easy to see that the centre of the distribution is 57 years, but the mode is lower, at 54 years.

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

It is also possible for there to be more than one mode for the same distribution of data, (bi-modal, or multi-modal). The presence of more than one mode can limit the ability of the mode in describing the centre or typical value of the distribution because a single value to describe the centre cannot be identified.

In some cases, particularly where the data are continuous, the distribution may have no mode at all (i.e. if all values are different).

In cases such as these, it may be better to consider using the median or mean, or group the data in to appropriate intervals, and find the modal class.

What is the median?

The median is the *middle value* in distribution when the values are arranged in ascending or descending order.

The median divides the distribution in half (there are 50% of observations on either side of the median value). In a distribution with an odd number of observations, the median value is the middle value.

Looking at the retirement age distribution (which has 11 observations), the median is the middle value, which is 57 years:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

When the distribution has an even number of observations, the median value is the mean of the two middle values. In the following distribution, the two middle values are 56 and 57, therefore the median equals 56.5 years:

52, 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

Advantage of the median:

The median is less affected by outliers and skewed data than the mean, and is usually the preferred measure of central tendency when the distribution is not symmetrical.

Limitation of the median:

The median cannot be identified for categorical nominal data, as it cannot be logically ordered.

What is the mean?

The mean is the sum of the value of each observation in a dataset divided by the number of observations. This is also known as the arithmetic average.

Looking at the retirement age distribution again:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

The mean is calculated by adding together all the values (54+54+54+55+56+57+57+58+58+60+60 = 623) and dividing by the number of observations (11) which equals 56.6 years.

Advantage of the mean:

The mean can be used for both continuous and discrete numeric data.

Limitations of the mean:

The mean cannot be calculated for categorical data, as the values cannot be summed.

As the mean includes every value in the distribution the mean is influenced by outliers and skewed distributions.

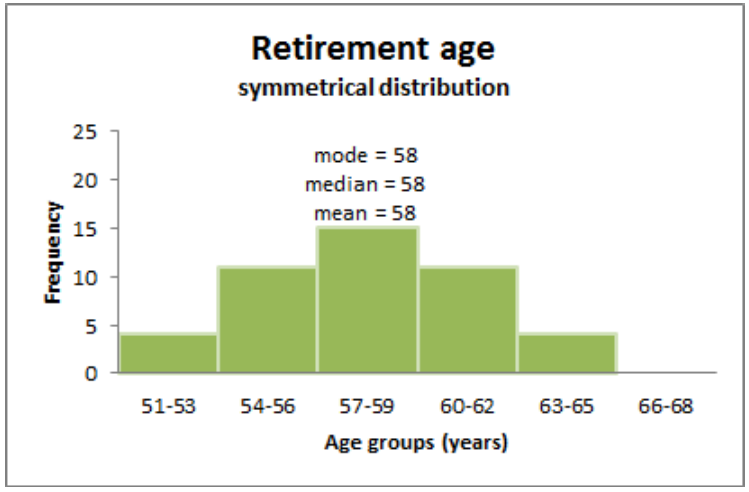
What else do I need to know about the mean?

The population mean is indicated by the Greek symbol μ (pronounced ‘mu’). When the mean is calculated on a distribution from a sample it is indicated by the symbol $\bar{x}$ (pronounced X-bar).

How does the shape of a distribution influence the Measures of Central Tendency?

Symmetrical distributions:

When a distribution is symmetrical, the mode, median and mean are all in the middle of the distribution. The following graph shows a larger retirement age dataset with a distribution which is symmetrical. The mode, median and mean all equal 58 years.

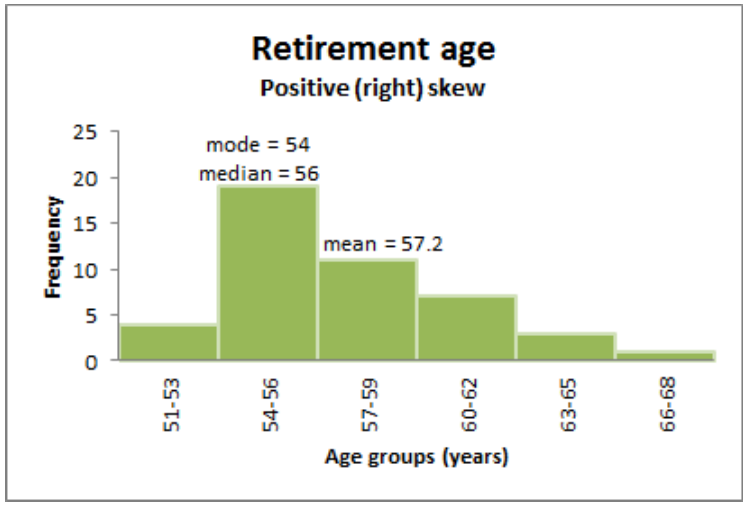


Skewed distributions:

When a distribution is skewed the mode remains the most commonly occurring value, the median remains the middle value in the distribution, but the mean is generally ‘pulled’ in the direction of the tails. In a skewed distribution, the median is often a preferred measure of central tendency, as the mean is not usually in the middle of the distribution.

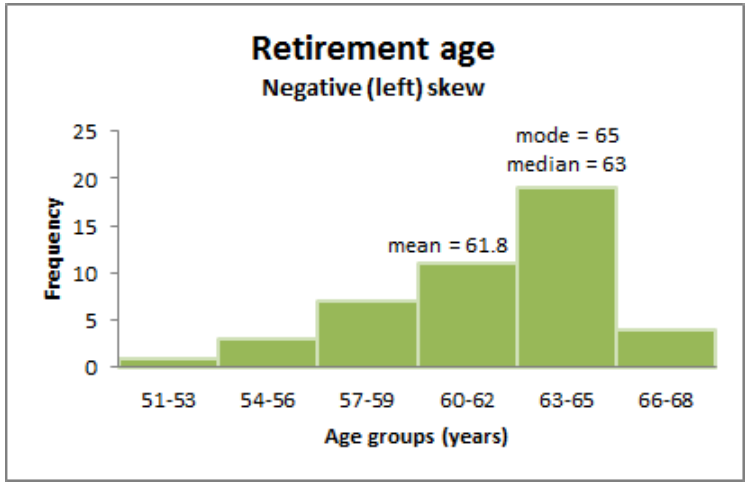
A distribution is said to be **positively or right skewed** when the tail on the right side of the distribution is longer than the left side. In a positively skewed distribution it is common for the mean to be ‘pulled’ toward the right tail of the distribution. Although there are exceptions to this rule, generally, most of the values, including the median value, tend to be less than the mean value.

The following graph shows a larger retirement age data set with a distribution which is right skewed. The data has been grouped into classes, as the variable being measured (retirement age) is continuous. The mode is 54 years, the modal class is 54-56 years, the median is 56 years and the mean is 57.2 years.



A distribution is said to be **negatively or left skewed** when the tail on the left side of the distribution is longer than the right side. In a negatively skewed distribution, it is common for the mean to be 'pulled' toward the left tail of the distribution. Although there are exceptions to this rule, generally, most of the values, including the median value, tend to be greater than the mean value.

The following graph shows a larger retirement age dataset with a distribution which left skewed. The mode is 65 years, the modal class is 63-65 years, the median is 63 years and the mean is 61.8 years.



How do outliers influence the measures of central tendency?

Outliers are extreme, or atypical data value(s) that are notably different from the rest of the data.

It is important to detect outliers within a distribution, because they can alter the results of the data analysis. The mean is more sensitive to the existence of outliers than the median or mode.

Consider the initial retirement age dataset again, with one difference; the last observation of 60 years has been replaced with a retirement age of 81 years. This value is much higher than the other values, and could be considered an outlier. However, it has not changed the middle of the distribution, and therefore the median value is still 57 years.

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 81

As the all values are included in the calculation of the mean, the outlier will influence the mean value.

(54+54+54+55+56+57+57+58+58+60+81 = 644), divided by 11 = 58.5 years

In this distribution the outlier value has increased the mean value.

Despite the existence of outliers in a distribution, the mean can still be an appropriate measure of central tendency, especially if the rest of the data is normally distributed. If the outlier is confirmed as a valid extreme value, it should not be removed from the dataset. Several common regression techniques can help reduce the influence of outliers on the mean value.

Further information:

ABS:
Education Services: Mean and Median Learning Tools

External links:
easycalculation.com - Mean, Median, Mode Calculator
calculatorsoup.com - Descriptive Statistics calculator
calculatorsoup.com - Mean Median Mode calculator

Return to Statistical Language Homepage

This page last updated 3 July 2013

Statistical Language - Measures of Spread

Measures of Spread

What are measures of spread?

Measures of spread describe how similar or varied the set of observed values are for a particular variable (data item). Measures of spread include the range, quartiles and the interquartile range, variance and standard deviation.

When can we measure spread?

The spread of the values can be measured for quantitative data, as the variables are numeric and can be arranged into a logical order with a low end value and a high end value.

Why do we measure spread?

Summarising the dataset can help us understand the data, especially when the dataset is large. As discussed in the Measures of Central Tendency page, the mode, median, and mean summarise the data into a single value that is typical or representative of all the values in the dataset, but this is only part of the 'picture' that summarises a dataset. Measures of spread summarise the data in a way that shows how scattered the values are and how much they differ from the mean value.

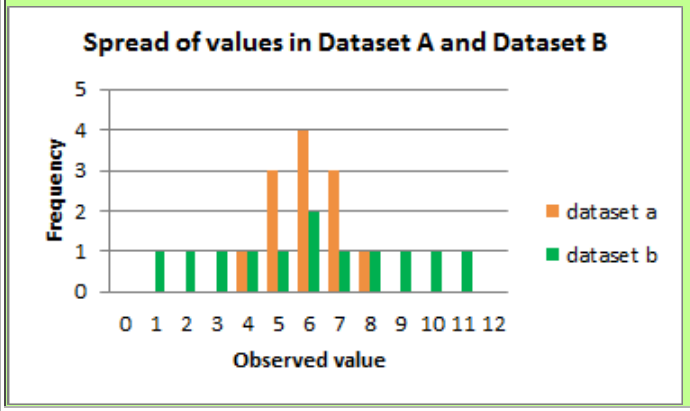
For example:

Dataset A	Dataset B
4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8	1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11

The mode (most frequent value), median (middle value*) and mean (arithmetic average) of both datasets is 6.
(*note, the median of an even numbered data set is calculated by taking the mean of the middle two observations).

If we just looked at the measures of central tendency, we may assume that the datasets are the same.

However, if we look at the spread of the values in the following graph, we can see that Dataset B is more dispersed than Dataset A. Used together, the measures of central tendency and measures of spread help us to better understand the data



What does each measure of spread tell us?

The range is the difference between the smallest value and the largest value in a dataset.

Calculating the Range

Dataset A
4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8

The range is 4, the difference between the highest value (8) and the lowest value (4).

Dataset B
1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11

The range is 10, the difference between the highest value (11) and the lowest value (1).

Dataset A

0	1	2	3	4	5	6	7	8	9	10	11	12	13

Dataset B

0	1	2	3	4	5	6	7	8	9	10	11	12	13

On a number line, you can see that the range of values for Dataset B is larger than Dataset A.

Quartiles divide an ordered dataset into four equal parts, and refer to the values of the point *between* the quarters. A dataset may also be divided into quintiles (five equal parts) or deciles (ten equal parts).

Quartiles							
25% of values	**Q1**	25% of values	**Q2**	25% of values	**Q3**	25% of values	

The **lower quartile (Q1)** is the point between the lowest 25% of values and the highest 75% of values. It is also called the **25th percentile**.

The **second quartile (Q2)** is the middle of the data set. It is also called the **50th percentile**, or the **median**.

The **upper quartile (Q3)** is the point between the lowest 75% and highest 25% of values. It is also called the **75th percentile**.

Calculating Quartiles

Dataset A

4	5	5	**Q1**	5	6	6	**Q2**	6	6	7	**Q3**	7	7	8

As the quartile point falls between two values, the mean (average) of those values is the quartile value:
Q1 = (5+5) / 2 = 5
Q2 = (6+6) / 2 = 6
Q3 = (7+7) / 2 = 7

Dataset B

1	2	3	**Q1**	4	5	6	**Q2**	6	7	8	**Q3**	9	10	11

As the quartile point falls between two values, the mean (average) of those values is the quartile value:
Q1 = (3+4) / 2 = **3.5**
Q2 = (6+6) / 2 = **6**
Q3 = (8+9) / 2 = **8.5**

The interquartile range (IQR) is the difference between the upper (Q3) and lower (Q1) quartiles, and describes the middle 50% of values when ordered from lowest to highest. The IQR is often seen as a better measure of spread than the range as it is not affected by outliers.

Interquartile Range						
25% of values	**Q1**	25% of values	**Q2**	25% of values	**Q3**	25% of values

Calculating the Interquartile Range

The IQR for Dataset A is = **2**
IQR = Q3 - Q1
= 7 - 5
= 2

The IQR for Dataset B is = **5**
IQR = Q3 - Q1
= 8.5 - 3.5
= 5

The variance and the standard deviation are measures of the spread of the data around the mean. They summarise how close each observed data value is to the mean value.

In datasets with a small spread all values are very close to the mean, resulting in a small variance and standard deviation. Where a dataset is more dispersed, values are spread further away from the mean, leading to a larger variance and standard deviation.

The smaller the variance and standard deviation, the more the mean value is indicative of the whole dataset. Therefore, if all values of a dataset are the same, the standard deviation and variance are zero.

The standard deviation of a normal distribution enables us to calculate confidence intervals. In a normal distribution, about 68% of the values are within one standard deviation either side of the mean and about 95% of the scores are within two standard deviations of the mean.

The population **Variance** σ^2 (pronounced *sigma squared*) of a discrete set of numbers is expressed by the following formula:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

where:

X_i represents the *ith* unit, starting from the first observation to the last

μ represents the population mean
N represents the number of units in the population

The **Variance** of a sample s^2 (pronounced *s squared*) is expressed by a slightly different formula:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

where:

x_i represents the *ith* unit, starting from the first observation to the last

$\bar{x}$ represents the sample mean
n represents the number of units in the sample

The **standard deviation** is the square root of the variance. The standard deviation for a population is represented by σ, and the standard deviation for a sample is represented by *S*.

Calculating the Population Variance σ^2 and Standard Deviation σ	
Dataset A	**Dataset B**
Calculate the population mean (μ) of Dataset A. (4 + 5 + 5 + 5 + 6 + 6 + 6 + 6 + 7 + 7 + 7 + 8) / 12 mean (μ) = 6 Calculate the deviation of the individual values from the mean by subtracting the mean from each value in the dataset $X_i - \mu$ = -2, -1, -1, -1, 0, 0, 0, 0, 1, 1, 1, 2 Square each individual deviation value $(X_i - \mu)^2$ = 4, 1, 1, 1, 0, 0, 0, 0, 1,1,1, 4 Calculate the mean of the squared deviation values $\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$ = (4 + 1 +1 +1 + 0 + 0 + 0 + 0 +1 +1 +1 + 4) / 12 **Variance σ^2= 1.17** Calculate the square root of the variance **Standard deviation σ = 1.08**	Calculate the population mean (μ) of Dataset B. (1 + 2 + 3 + 4 + 5 + 6 + 6 + 7 + 8 + 9 + 10 + 11) / 12 mean (μ) = 6 Calculate the deviation of the individual values from the mean by subtracting the mean from each value in the dataset $X_i - \mu$ = -5, -4, -3, -2, -1, 0, 0, 1, 2, 3, 4, 5, Square each individual deviation value $(X_i - \mu)^2$ = 25, 16, 9, 4, 1, 0, 0, 1, 4, 9, 16, 25 Calculate the mean of the squared deviation values $\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$ = (25 + 16 + 9 + 4 + 1 + 0 + 0 + 1 + 4 + 9 + 16 + 25) / 12 **Variance σ^2 = 9.17** Calculate the square root of the variance **Standard deviation σ = 3.03**

The larger Variance and Standard Deviation in Dataset B further demonstrates that Dataset B is more dispersed than Dataset A.

Further information:

External links:

easycalculation.com - Standard Deviation calculator
easycalculation.com - Five Number Summary calculator

Return to Statistical Language Homepage

This page last updated 4 July 2013

Statistical Language - Types of Error

Types of Error

What is error?

Error (statistical error) describes the difference between a value obtained from a data collection process and the 'true' value for the population. The greater the error, the less representative the data are of the population.

Data can be affected by two types of error: sampling error and non-sampling error.

What is sampling error?

Sampling error occurs solely as a result of using a sample from a population, rather than conducting a census (complete enumeration) of the population. It refers to the difference between an estimate for a population based on data from a sample and the 'true' value for that population which would result if a census were taken. Sampling errors do not occur in a census, as the census values are based on the entire population.

Sampling error can occur when:

- the proportions of different characteristics within the sample are not similar to the proportions of the characteristics for the whole population (i.e. if we are taking a sample of men and women and we know that 51% of the total population are women and 49% are men, then we should aim to have similar proportions in our sample);
- the sample is too small to accurately represent the population; and
- the sampling method is not random.

Sampling error can be measured and controlled in random samples where each unit has a chance of selection, and that chance can be calculated. In general, increasing the sample size will reduce the sample error.

What is non-sampling error?

Non-sampling error is caused by factors other than those related to sample selection. It refers to the presence of any factor, whether systemic or random, that results in the data values not accurately reflecting the 'true' value for the population.

Non-sampling error can occur at any stage of a census or sample study, and are not easily identified or quantified.

Non-sampling error can include (but is not limited to):

- **Coverage error:** this occurs when a unit in the sample is incorrectly excluded or included, or is duplicated in the sample (e.g. a field interviewer fails to interview a selected household or some people in a household).
- **Non-response error:** this refers to the failure to obtain a response from some unit because of absence, non-contact, refusal, or some other reason. Non-response can be complete non-response (i.e. no data has been obtained at all from a selected unit) or partial non-response (i.e. the answers to some questions have not been provided by a selected unit).
- **Response error:** this refers to a type of error caused by respondents intentionally or accidentally providing inaccurate responses. This occurs when concepts, questions or instructions are not clearly understood by the respondent; when there are high levels of respondent burden and memory recall required; and because some questions can result in a tendency to answer in a socially desirable way (giving a response which they feel is more acceptable rather than being an accurate response).
- **Interviewer error:** this occurs when interviewers incorrectly record information; are not neutral or objective; influence the respondent to answer in a particular way; or assume responses based on appearance or other characteristics.
- **Processing error:** this refers to errors that occur in the process of data collection, data entry, coding, editing and output.

Why does error matter?

The greater the error the less reliable are the results of the study. A credible data source will have measures in place throughout the data collection process to minimise the amount of error, and will also be transparent about the size of the expected error so that users can decide whether the data are 'fit for purpose'.

Examples of question wording which may contribute to non-sampling error.

Memory recall:
"How many kilometres did you travel in July last year?"

Socially desirable questions:
"Do you regularly recycle your waste paper and plastics?"

Under reporting:
"How many glasses of alcohol do you drink per week?"

Over reporting:
"How much did you win from gambling last week?"

Leading question:
"Do you think the government is doing enough to stop the increase in violent crime on our streets?"

Double-barrelled question:
"Are you happy with the price of, and services offered by, your gym membership?"

Recommended: Read Measures of Error next

Further information:

External links:

- Errors in Statistical data

Return to Statistical Language Homepage

This page last updated 3 July 2013

Statistical Language - Measures of Error

Measures of Error

Recommended: First read Types of Error

Why do we measure error?

Error is expected in a data collection process, particularly if the data is obtained from a sample survey. Although non-sampling error is difficult to measure, sampling error can be measured to give an indication of the accuracy of any estimate value for the population. This assists users to make informed decisions about whether the statistics are suited to their needs.

How do we measure error?

Two common measures of error include the standard error and the relative standard error.

Standard Error (SE) is a measure of the variation between any estimated population value that is based on a sample rather than true value for the population. As the standard error of an estimated value generally increases with the size of the estimate, a large standard error may not necessarily result in an unreliable estimate. Therefore it is often better to compare the error in relation to the size of the estimate.

Relative Standard Error (RSE) is the standard error expressed as a proportion of an estimated value. It is usually displayed as a percentage. RSEs are a useful measure as they provide an indication of the relative size of the error likely to have occurred due to sampling. A high RSE indicates less confidence that an estimated value is close to the true population value.

Where published statistics contain an indication of the RSEs they can be used to compare statistics from different studies of the same population.

What can measures of error tell us?

The standard error can be used to construct a confidence interval.

A confidence interval is a range in which it is estimated the true population value lies. Confidence intervals of different sizes can be created to represent different levels of confidence that the true population value will lie within a particular range. A common confidence interval used in statistics is the 95% confidence interval. In a 'normal distribution', the 95% confidence interval is measured by two standard errors either side of the estimate.

Further information:

ABS:
What is a Standard Error and Relative Standard Error?

External link:
easycalculation.com - Population Confidence Interval calculator

Return to Statistical Language Homepage

This page last updated 3 July 2013

Statistical Language - What are Statistics?

Statistical Language

What are Statistics?

This animation explains what are statistics, if you are unable to access the video a Transcript (.doc 32kb) *has been provided. The animation requires* Adobe Flash Player *to run. The animation contains no audio.*

How can statistics help us?

A statistic is a value that has been produced from a data collection, such as a summary measure, an estimate or projection. Statistical information is data that has been organised to serve a useful purpose.

Statistics is also a term that refers to the practice of collecting, analysing, interpreting and communicating data. It is the science of interacting with data.

How can statistics help us?

Statistics represent a common method of presenting information helping us to understand what the data are telling us.

Descriptive (or summary) statistics summarise the raw data and allow data users to interpret a dataset more easily.

Descriptive statistics can describe the shape, centre and spread of a dataset..

Inferential statistics are used to infer conclusions about a population from a sample of that population. Inferential statistics are the result of techniques that use the data collected from a sample to make generalisations about the whole population from which the sample was taken.

Inferential statistics include estimation, and hypothesis testing.

Why are statistics important?

Statistics form an evidence base for decision making and help us identify issues and opportunities, develop options and recommendations, monitor progress, evaluate outcomes, and understand the world.

Statistics that are published by government agencies, such as the ABS, or other official organisations are called “official statistics”.

Principle 1 of the United Nations Statistical Commission's Fundamental Principles of Official Statistics states that:

'Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honor citizens' entitlement to public information.'

Further information:

ABS:
1500.0 - A guide for using statistics for evidence based policy

External links:
United Nations Statistical Commission's Fundamental Principles of Official Statistics

---

Return to Statistical Language Homepage

This page last updated 3 July 2013

Statistical Language - What are Standards?

What are Standards?

What are statistical standards?

A statistical standard is a set of rules used to standardise the way data are collected and statistics are produced. They provide information about data collected on a particular topic that assists in the understanding and interpretation of that data.

Statistical standards are the approved versions of how to:

- define the underlying concept
- define the specific variables of interest
- collect data
  - coding structure
  - statistical units
  - recommended question modules
- process the data
  - classification
  - standard editing
- present the data
  - output categories
- interpret the data

What are classifications?

Classifications are used to collect and organise information into categories with other similar pieces of information. They are an important part of any standard.

Classifications are used in most parts of the statistical cycle including:

- data collection
- data processing
- data presentation
- data analysis

Classifications should be exhaustive, and mutually exclusive.

For example, Australia, France, Japan, and Vanuatu are some of the categories from the ‘Standard Australian Classification of Countries’ (SACC). This classification is used in the ‘Country of Birth Standard’ where each response for a person's country of birth will belong in one and only one category within the classification.

Why do we need standards in statistics?

Standards are used to ensure that data about the same characteristic are collected and communicated in the same way every time. This enables data from different sources to be compared on a consistent basis and enables meaningful comparisons to be made over time.

There are four key advantages from having widespread use of approved statistical standards:

1. ensure the quality of statistical outputs
2. creates a meaningful statistical picture of society and economy
3. reduces costs
4. improves transparency

How are statistical standards used?

Standards are used:

- internally in the Australian Bureau of Statistics to ensure data are produced to a consistent quality and in a consistent manner over time and across collections
- nationally by those producing statistics to assist with the integration of data from various sources
- internationally to comply with international reporting obligations and encourage data comparability between countries.

Comparability is the ability to validly compare statistics that have been collected over time, or from different sources.

For example, in the standard definitions used for ABS labour force collections, a person must be actively seeking employment and available to start work to be classified as unemployed. If the definition of unemployed is changed to include people who were not actively seeking work and/or not available to start work, the two datasets would not be comparable as the data has not been collected using common definitions. It would be difficult for users to determine whether a change in the reported unemployment rate was due to a change in definitions used or a reflection of what is happening in the labour market.

Further information

ABS:
Methods & Standards

External links:
National Statistical Service (NSS) - Standard Classifications
Standards Australia

Return to Statistical Language Homepage

This page last updated 3 July 2013

Statistical Language - What is Metadata?

Statistical Language

What is Metadata?

This animation explains the concept of metadata. If you are unable to access the video a Transcript (.doc 27kb) *has been provided. The animation requires* Adobe Flash Player *to run. The animation contains no audio.*

What is metadata?

Metadata is the information that defines and describes data.

It is often referred to as *data about data* or *information about data* because it provides data users with information about the purpose, processes, and methods involved in the data collection.

How is metadata presented?

Metadata provides information about all aspects of data collection; from design through to communication.

For example, metadata may appear alongside the data in the form of graph labels and footnotes, or may be compiled as explanatory notes that contain information such as a definition and description of the population, the source of the data, and the methodology used, to assist with the interpretation of the data.

Why is metadata important?

The processes used to collect data and produce statistics may influence the suitability of the information for different statistical purposes.

Metadata provides information to enable the user to make an informed decision about whether the data are fit for the required purpose.

Further information

External links:
About Metadata
OECD Definition of Statistical Metadata

Return to Statistical Language Homepage

This page last updated 3 July 2013

Statistical Language - Data Visualisation

Data Visualisation

What is data visualisation?

Data visualisation involves the visual presentation of data to communicate the stories contained in the dataset.

Data visualisation can communicate complex information in a way that is easier to interpret by turning data into visually engaging images and 'stories'.

How can data be visualised?

Data can be communicated visually through:

Static visualisation - the use of graphs and charts which provides a visual snapshot of the data.

Dynamic visualisation - the use of animations which emphasise key information and show movements in the data.

Interactive visualisation - data users are able to change the graphics so as to view different variables. This provides opportunities for users to become active data explorers with the freedom to customise what they see, look deeper into specific areas of the data, or use motion to track patterns over time and space.

How can data visualisation be used?

Data visualisation not only communicates data in an easy to understand way, it can also be used as a tool for data analysis as patterns, trends, relationships between variables, and the distribution of the dataset can be more apparent than when presented as numbers in a table.

The use of data visualisation for analysis can assist with exploring the data and guiding the direction of further data investigation. It is also possible to reduce complex datasets (or integrate multiple datasets) to reveal specific characteristics of interest, explore changes over time, or investigate relationships between variables.

When is data visualisation suitable?

Data visualisation is beneficial when the user needs an overview of a dataset rather than the specific values contained in the dataset. The visualisation will highlight the stories in the data, or focus on selected data from within the dataset for a specific purpose.

Data visualisation links:

ABS:
Visit Interact with our data on the ABS website to explore the visualisation of ABS data.

External links:
TED (Technology, Entertainment, Design)
Gapminder
Statistics Netherlands interactive info graphs

Return to Statistical Language Homepage

This page last updated 3 July 2013

Statistical Language - Time Series Data

Statistical Language

Time Series Data

What is a time series?

A time series is a collection of observations of well-defined data items obtained through repeated measurements over time.

For example, measuring the level of unemployment each month of the year would comprise a time series. This is because employment and unemployment are well defined, and consistently measured at equally spaced intervals. Data collected irregularly or only once are not time series.

What does a time series show?

A times series allows you to identify change within a population over time. A time series can also show the impact of cyclical, seasonal and irregular events on the data item being measured.

Time series can be classified into two different types: stock and flow.

A stock series is a measure of certain attributes at a point in time and can be thought of as "stock takes". For example, the annual ABS *Prisoners in Australia* collection is a stock measure because it is a count of the number of persons in custody who were the legal responsibility of adult corrective services agencies on the night of 30 June each year.

A flow series is a series which is a measure of activity over a given period. For example, the quarterly ABS *Corrective Services, Australia* collection is a flow measure as it provides prisoner counts taken on each day of the month which are summed and divided by the number of days in that month to determine the mean (average) daily prisoner number for that month.

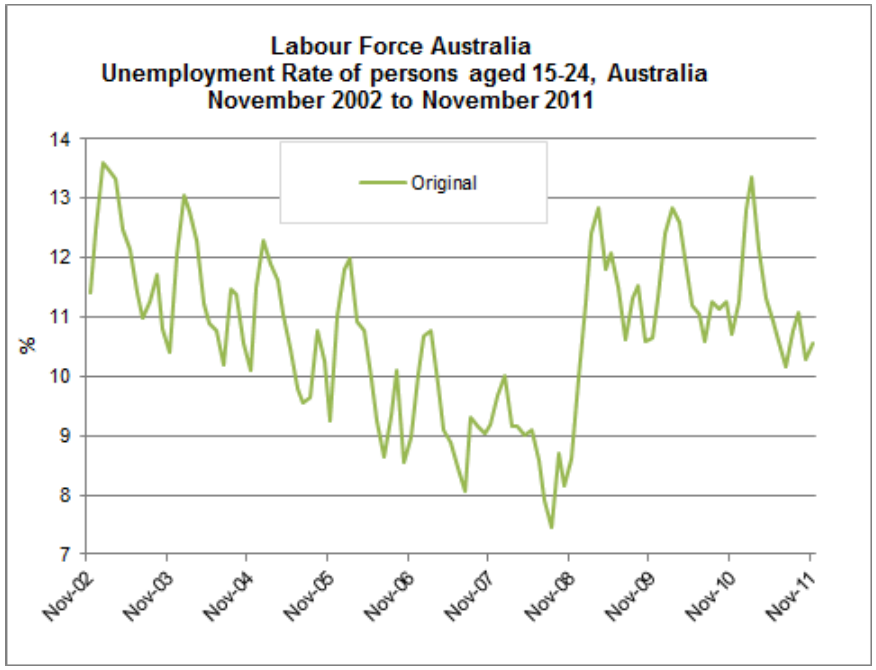
How can a time series be analysed?

An original time series shows the actual movements in the data over time. An original series includes any movements due to cyclical, seasonal and irregular events.

A cyclical effect is any regular fluctuation in daily, weekly, monthly or annual data. For example, the number of commuters using public transport has regular peaks and troughs during each day of the week, depending on the time of day.

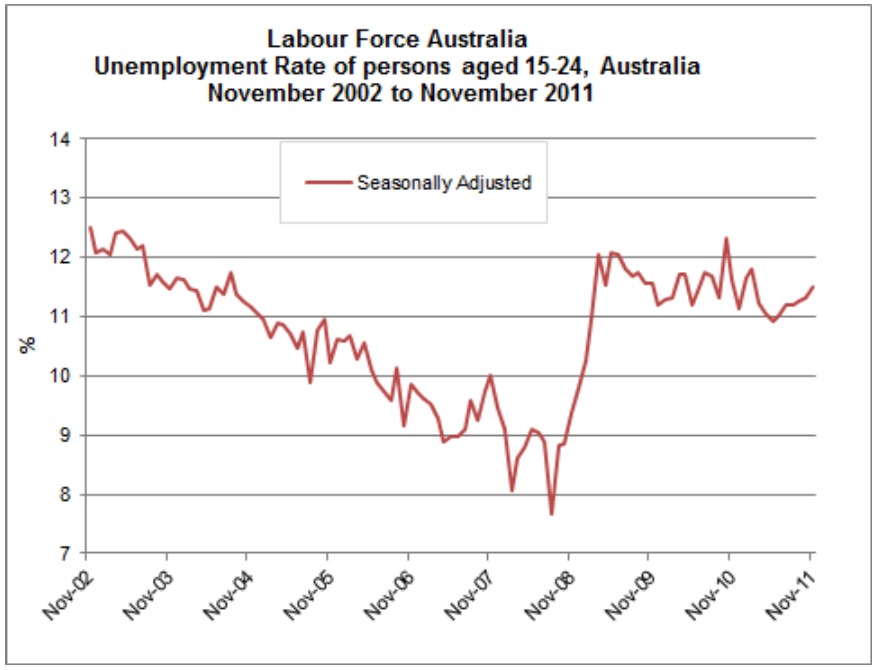
A seasonal effect is any variation in data due to calendar related effects which occur systematically at specific seasonal frequencies every year. For example, in Australia employment increases over the Christmas/New Year period, or fruit and vegetable prices can vary depending on whether or not they are 'in-season'.

An irregular effect is any movement that occurred at a specific point in time, but is unrelated to a season or cycle. For example, a natural disaster, the introduction of legislation, or a one-off major cultural or sporting event.



A seasonally adjusted series involves estimating and removing the cyclical and seasonal effects from the original data. Seasonally adjusting a time series is useful if you wish to understand the underlying patterns of change or movement in a population, without the impact of the seasonal or cyclical effects.

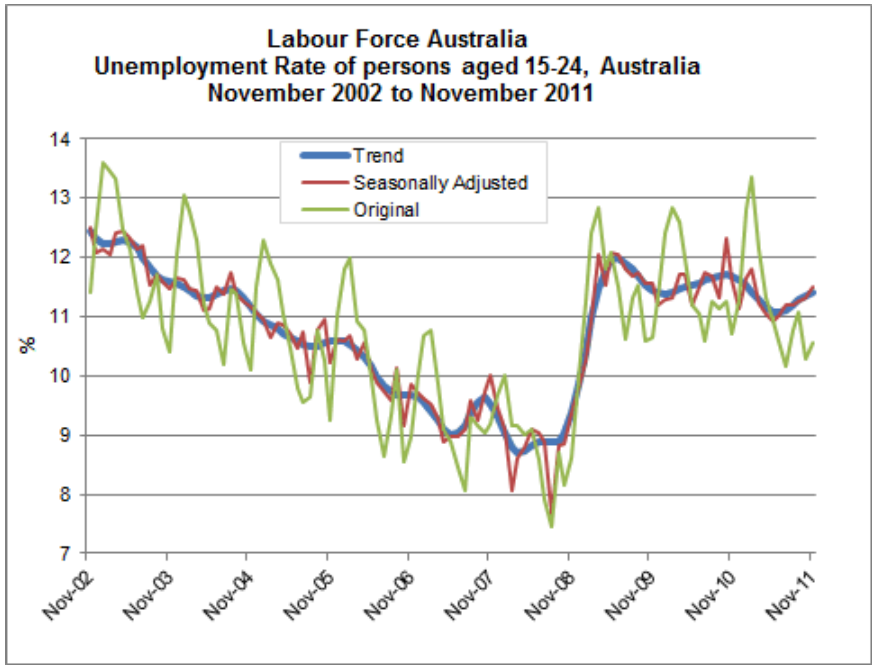
For example, employment and unemployment are often seasonally adjusted so that the actual change in employment and unemployment levels can be seen, without the impact of periods of peak employment such as Christmas/New Year when a large number of casual workers are temporarily employed.



A trend series is a seasonally adjusted series that has been further adjusted to remove irregular effects and 'smooth' out the series to show the overall 'trend' of the data over time.

For example, the trend is often used when analysing economic indicators such as employment and unemployment levels.

A graph combining the three series shows an example of the influence the cyclical, seasonal, and irregular effects can have on values within a dataset.



Further information:

ABS:
3222.0 - Population Projections, Australia, 2006 to 2101
Animated Historical Population Chart
Tourist Accommodation - Room Occupancy Rates Chart
1349.0 - Information Paper: A Guide to Interpreting Time Series - Monitoring Trends
1346.0.55.001 - Information paper: An Introductory Course on Time Series Analysis - Electronic Delivery

External link:
The Joy of Stats (Gapminder)

Return to Statistical Language Homepage

This page last updated 3 July 2013

Statistical Language - Estimate and Projection

Statistical Language

Estimate and Projection

What is an estimate?

An estimate is a value that is inferred for a population based on data collected from a sample of units from that population. Estimation is a technique that systematically adjusts the sample data to determine an estimated value for the population.

For example, if our sample data shows that 51% of the sample are female, then the population value will be estimated to be 51% (as estimation is based on the assumption that the sample is representative of the population).

An estimate is not a guess, it is a value based on sampled data which has been adjusted using statistical estimation procedures.

What is a projection?

A projection indicates what the future changes in a population would be if the assumptions about future trends actually occur. These assumptions are often based on patterns of change which have previously occurred.

For example: Data collected about the total number of store locations for a retail chain over three years show an increase from 8 stores in first year, to 12 stores in the second year, to 18 stores in the third year. It could therefore be projected that if the chain continues to expand following the same pattern of increasing by half (50%) each year there will be 27 stores after the fourth year.

A projection is not making a prediction or forecast about what is going to happen, it is indicating what would happen if the assumptions which underpin the projection actually occur.

Comparison of Projections and Forecasts

Type of Information	The Difference	Nature of Assumptions
Projections indicate what future values for the population would be if the assumed patterns of change were to occur. They are not a prediction that the population will change in this manner.	While both involve analysis of data, the key difference between a forecast and a projection is the nature of the assertion in	A **projection** simply indicates a future value for the population if the set of underlying assumptions occur.

Forecasts speculate future values for the population with a certain level of confidence, based on current and past values as an expectation (prediction) of what will happen.	relation to the assumptions occurring.	In a **forecast**, the assumptions represent expectations of actual future events.

How do estimates and projections differ?

An estimate is a statistic about a whole population for a previous reference period which is based on data from a sample of the population, whereas a projection is a statistic indicating what a value would be if the assumptions about future trends hold true (often drawing upon past movements in a population as a guide for the assumptions).

Further information

ABS:
Animated Population Pyramid
3222.0 - Population Projections, Australia, 2006 to 2101
3228.0.55.001 - Population Estimates: Concepts, Sources and Methods
Health - If Australia were only 100 people...

Return to Statistical Language Homepage

This page last updated 3 July 2013

Statistical Language - Correlation and Causation

Correlation and Causation

This animation explains the concept of correlation and causation. If you are unable to access the video a Transcript (.doc 26kb) *has been provided. The animation requires* Adobe Flash Player *to run. The animation contains no audio.*

What are correlation and causation and how are they different?

Two or more variables considered to be related, in a statistical context, if their values change so that as the value of one variable increases or decreases so does the value of the other variable (although it may be in the opposite direction).

For example, for the two variables "hours worked" and "income earned" there is a relationship between the two if the increase in hours worked is associated with an increase in income earned. If we consider the two variables "price" and "purchasing power", as the price of goods increases a person's ability to buy these goods decreases (assuming a constant income).

Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables. A correlation between variables, however, does not automatically mean that the change in one variable is the cause of the change in the values of the other variable.

Causation indicates that one event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events. This is also referred to as cause and effect.

Theoretically, the difference between the two types of relationships are easy to identify — an action or occurrence can *cause* another (e.g. smoking causes an increase in the risk of developing lung cancer), or it can *correlate* with another (e.g. smoking is correlated with alcoholism, but it does not cause alcoholism). In practice, however, it remains difficult to clearly establish cause and effect, compared with establishing correlation.

Why are correlation and causation important?

The objective of much research or scientific analysis is to identify the extent to which one variable relates to another variable. For example:

- Is there a relationship between a person's education level and their health?
- Is pet ownership associated with living longer?
- Did a company's marketing campaign increase their product sales?

These and other questions are exploring whether a correlation exists between the two variables, and if there is a correlation then this may guide further research into investigating whether one action causes the other. By understanding correlation and causality, it allows for policies and programs that aim to bring about a desired outcome to be better targeted.

How is correlation measured?

For two variables, a statistical correlation is measured by the use of a Correlation Coefficient, represented by the symbol (r), which is a single number that describes the degree of relationship between two variables.

The coefficient's numerical value ranges from +1.0 to –1.0, which provides an indication of the strength and direction of the relationship.

If the correlation coefficient has a negative value (below 0) it indicates a negative relationship between the variables. This means that the variables move in opposite directions (ie when

one increases the other decreases, or when one decreases the other increases).

If the correlation coefficient has a positive value (above 0) it indicates a positive relationship between the variables meaning that both variables move in tandem, i.e. as one variable decreases the other also decreases, or when one variable increases the other also increases.

Where the correlation coefficient is 0 this indicates there is no relationship between the variables (one variable can remain constant while the other increases or decreases).

While the correlation coefficient is a useful measure, it has its limitations:

Correlation coefficients are usually associated with measuring a linear relationship.
For example, if you compare hours worked and income earned for a tradesperson who charges an hourly rate for their work, there is a linear (or straight line) relationship since with each additional hour worked the income will increase by a consistent amount.

If, however, the tradesperson charges based on an initial call out fee and an hourly fee which progressively decreases the longer the job goes for, the relationship between hours worked and income would be non-linear, where the correlation coefficient may be closer to 0.

Care is needed when interpreting the value of 'r'. It is possible to find correlations between many variables, however the relationships can be due to other factors and have nothing to do with the two variables being considered.
For example, sales of ice creams and the sales of sunscreen can increase and decrease across a year in a systematic manner, but it would be a relationship that would be due to the effects of the season (ie hotter weather sees an increase in people wearing sunscreen as well as eating ice cream) rather than due to any direct relationship between sales of sunscreen and ice cream.

The correlation coefficient should not be used to say anything about cause and effect relationship. By examining the value of 'r', we may conclude that two variables are related, but that 'r' value does not tell us if one variable was the cause of the change in the other.

How can causation be established?

Causality is the area of statistics that is commonly misunderstood and misused by people in the mistaken belief that because the data shows a correlation that there is necessarily an underlying causal relationship .

The use of a controlled study is the most effective way of establishing causality between variables. In a controlled study, the sample or population is split in two, with both groups being comparable in almost every way. The two groups then receive different treatments, and the outcomes of each group are assessed.

For example, in medical research, one group may receive a placebo while the other group is given a new type of medication. If the two groups have noticeably different outcomes, the different experiences may have caused the different outcomes.

Due to ethical reasons, there are limits to the use of controlled studies; it would not be appropriate to use two comparable groups and have one of them undergo a harmful activity while the other does not. To overcome this situation, observational studies are often used to investigate correlation and causation for the population of interest. The studies can look at the groups' behaviours and outcomes and observe any changes over time.

The objective of these studies is to provide statistical information to add to the other sources of information that would be required for the process of establishing whether or not causality exists between two variables.

Further information

ABS:
1500.0 - A guide for using statistics for evidence based policy
Literacy Stats: Using ABS Statistics: Telling the right story

Return to Statistical Language Homepage

This page last updated 3 July 2013

Statistical Language - Confidentiality

Confidentiality

This animation explains the concept of confidentiality. If you are unable to access the video a Transcript (.doc 27kb) *has been provided. The animation requires* Adobe Flash Player *to run. There is no audio in this animation.*

What is confidentiality?

Confidentiality refers to the obligation of organisations that collect information to ensure that no person or organisation is likely to be identified from any data released.

Why is confidentiality necessary?

Organisations that collect data depend on the goodwill and cooperation of the community, businesses and other organisations to provide the information. By protecting the confidentiality of the information provided, organisations that collect data help maintain the trust and goodwill of providers, and are better able to collect the required information. Maintaining public trust helps achieve a higher response to data collections and results in better quality data.

There are also legal obligations which must be met in relation to the collection, management, use and dissemination of information. In Australia this requirement is recognised in the Commonwealth Privacy Act (1988) and various state and territory privacy legislation. It is also reflected in legislation, procedures and protocols in relation to specific government-activities where information is collected. Examples include the Social Security (Administration) Act 1999, the Taxation Administration Act 1953 and the Census and Statistics Act (1905). Penalties apply if the secrecy provisions set out in these Acts are breached.

At the international level, the United Nations Statistical Commission identifies confidentiality as one of the *Fundamental Principles of Official Statistics*, with principle 6 stating: *Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.*

How is data kept confidential?

Organisations that collect data protect the secrecy of information by implementing policies and procedures that address all aspects of data protection.

They do this by ensuring identifiable information about individuals and organisations;

- is not released publicly;
- is available to authorised people on a need to know basis only;
- cannot be derived from disseminated data;
- and is maintained and accessed securely.

To avoid the disclosure of confidential information where an individual person or organisation could be identified in a dataset, either directly or indirectly, the data are confidentialised. This involves removing or altering information, or collapsing detail, to ensure that no person or organisation is likely to be identified in the data. There are various methods used to protect the identity of individuals and organisations while at the same time maximising the usefulness of the data for statistical and research purposes.

Further information:

ABS:
2011 Census - Privacy and Confidentiality
ABS Web Site Privacy Statement
Survey Participant Information - How The ABS Keeps Your Information Confidential

External Links:
National Statistical Service (NSS) - Confidentiality Information Series
Fundamental Principles of Official Statistics

Return to Statistical Language Homepage

This page last updated 18 June 2013

Statistical Language - Statistical Language Glossary

Statistical Language

Statistical Language Glossary

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

A

Absolute frequency

The absolute frequency describes the number of times a particular value for a variable (data item) has been observed to occur.

See: Describing Frequencies

Administrative data

Administrative data are collected as part of the day to day processes and record keeping of organisations.

See: Data Sources

B

Bar chart

A bar chart is a type of graph in which each column (plotted either vertically or horizontally) represents a categorical variable or a discrete ungrouped numeric variable.

See: Frequency Distribution

C

Categorical variable

Categorical variables have values that describe a 'quality' or 'characteristic' of a data unit, like 'what type' or 'which category'.

See: What are Variables?

Causation

Causation indicates that one event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events. This is also referred to as cause and effect.

See: Correlation and Causation

Census (complete enumeration)

A census is a study of every unit, everyone or everything, in a population.

See: Census and Sample

Classifications

Classifications are used to collect and organise information into categories with other similar pieces of information.

See: What are Standards?

Class interval

A class interval is a range of data values. Each class interval has a lower and upper limit and contains all observations with values in that range. Class intervals cannot overlap with one another. For example 0 - 4, 5 - 8, 9 - 12.

Cohort

A cohort is a group of data units sharing a common experience or characteristic.

Comparability

Comparability is the ability to validly compare statistics that have been collected over time, or from different sources.

See: What are Standards?

Confidence interval

A confidence interval is a range in which it is estimated the true population value lies.

See: Measures of Error

Confidentiality

Confidentiality refers to the obligation of organisations that collect information to ensure that no person or organisation is likely to be identified from any data released.

See: Confidentiality

Continuous variable

A continuous variable is a numeric variable. Observations can take any value between a certain set of real numbers.

See: What are Variables?

Correlation

Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables.

See: Correlation and Causation

Coverage

The coverage is the actual population of units within the scope of a data collection about which data can actually be collected. As it is not always possible to collect data from units in the population of interest, units may be in scope but not in coverage.

See also: Scope

Cyclical effect

A cyclical effect is any regular fluctuation in daily, weekly, monthly or annual data.

See: Time Series Data

D

Data

Data are measurements or observations that are collected as a source of information.

See: What are Data?

Data item (or variable)

A data item is a characteristic (or attribute) of a data unit which is measured or counted, such as height, country of birth, or income.

See: What are Data?

Dataset

A dataset is a complete collection of all observations.

See: What are Data?

Data unit

A data unit is one entity (such as a person or business) in the population being studied, about which data are collected.

See: What are Data?

Data visualisation

Data visualisation involves the visual presentation of data to communicate the stories contained in the dataset.

See: Data Visualisation

Descriptive (or summary) statistics

Descriptive statistics summarise the raw data and allow data users to interpret a dataset more easily.

See: What are Statistics?

Discrete variable

A discrete variable is a numeric variable. Observations can take a value based on a count from a set of distinct whole values.

See: What are Variables?

E

Error (Statistical error)

Statistical error **describes the difference between a value obtained from a data collection process and the 'true' value for the population.**

See: Types of Error

Estimate

An estimate is a value that is inferred for a population based on data collected from a sample of units from that population.

See: Estimate and Projection

F

Flow series

A flow series is a series which is a measure of activity over a given period.

See: Time Series Data

Frequency

The frequency is the number of times a particular value for a variable (data item) has been observed to occur.

See: Describing Frequencies

Frequency distribution

Frequency distributions are visual displays that organise and present frequency counts so that the information can be interpreted more easily.

See: Frequency Distribution

H

Histogram

A histogram is a type of graph in which each column represents a numeric variable, in particular that which is continuous and/or grouped.

See: Frequency Distribution

I

Index number

An index number is a ratio measuring the value of a data item at one time in relation to its value at a base period. Index numbers measure change without giving the actual numerical value of the data item.

Inferential statistics

Inferential statistics are used to infer conclusions about a population from a sample of that population.

See: What are Statistics?

Interquartile range (IQR)

The interquartile range (IQR) is the difference between the upper (Q3) and lower (Q1) quartiles, and describes the middle 50% of values when ordered from lowest to highest.

See: Measures of Spread

Irregular effect

An irregular effect is any movement that occurred at a specific point in time, but is unrelated to a season or cycle.

See: Time Series Data

M

Mean

The mean is the sum of the value of each observation in a dataset divided by the number of observations. This is also known as the arithmetic average.

See: Measures of Central Tendency

Measures of central tendency (centre or central location)

A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.

See: Measures of Central Tendency

Measures of shape

Measures of shape **describe the distribution (or pattern) of the data within a dataset.**

See: Measures of Shape

Measures of spread

Measures of spread **describe how similar or varied the set of observed values are for a particular variable (data item).**

See: Measures of Spread

Median

The median is the *middle value* in distribution when the values are arranged in ascending or descending order.

See: Measures of Central Tendency

Metadata

Metadata **is the information that defines and describes data.**

See: What is Metadata?

Mode

The mode is the most commonly occurring value in a distribution.

See: Measures of Central Tendency

N

Nominal variable

A nominal variable is a categorical variable. Observations can take a value that is not able to be organised in a logical sequence.

See: What are Variables?

Non-random (non-probability) sample

In a non-random (or non-probability) sample some units of the population have no chance of selection, the selection is non-random, or the probability of their selection can not be determined.

See: Census and Sample

Non-sampling error

Non-sampling error is caused by factors other than those related to sample selection.

See: Types of Error

Normal distribution

A normal distribution is a true symmetric distribution of observed values.

See: Measures of Shape

Numeric variable

Numeric variables have values that describe a measurable quantity as a number, like 'how many' or 'how much'.

See: What are Variables?

O

Observation

An observation is an occurrence of a specific data item that is recorded about a data unit.

See: What are Data?

Ordinal variable

An ordinal variable is a categorical variable. Observations can take a value that can be logically ordered or ranked.

See: What are Variables?

Original time series

An original time series shows the actual movements in the data over time.

See: Time Series Data

Outlier

Outliers are extreme, or atypical data value(s) that are notably different from the rest of the data.

See: Measures of Central Tendency

P

Percentage

A percentage expresses a value for a variable in relation to a whole population as a fraction of one hundred.

See: Describing Frequencies

Population

A population is any complete group with at least one characteristic in common.

See: What is a Population?

Projection

A projection indicates what the future changes in a population would be if the assumptions about future trends actually occur.

See: Estimate and Projection

Proportion

A proportion describes the share of one value for a variable in relation to a whole.

See: Describing Frequencies

Q

Qualitative data

Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.

See: Quantitative and Qualitative Data

Quantitative data

Quantitative data are measures of values or counts and are expressed as numbers.

See: Quantitative and Qualitative Data

Quartiles

Quartiles divide an ordered dataset into four equal parts, and refer to the values of the point *between* the quarters. A dataset may also be divided into quintiles (five equal parts) or deciles (ten equal parts).

See: Measures of Spread

R

Random (probability) sample

In a random (or probability) sample each unit in the population has a chance of being selected, and this probability can be accurately determined.

See: Census and Sample

Range

The range is the difference between the smallest value and the largest value in a dataset.

See: Measures of Spread

Rate

A rate is a measurement of one value for a variable in relation to another measured quantity.

See: Describing Frequencies

Ratio

A ratio compares the frequency of one value for a variable with another value for the variable.

See: Describing Frequencies

Relative frequency

A relative frequency describes the number of times a particular value for a variable (data item) has been observed to occur in relation to the total number of values for that variable.

See: Describing Frequencies

Relative standard error (RSE)

The relative standard error (RSE) is the standard error expressed as a proportion of an estimated value.

See: Measures of Error

Respondent

A respondent provides data about oneself as a unit, or as a representative of another unit in a population.

See: Data Sources

S

Sample (partial enumeration)

A sample is a subset of units in a population, selected to represent all units in a population of interest.

See: Census and Sample

Sampling error

Sampling error occurs solely as a result of using a sample from a population, rather than conducting a census (complete enumeration) of the population.

See: Types of Error

Scope

The scope is the set of units that comprise the population of interest (target population) about which data are being collected.

See also: Coverage

Seasonal effect

A seasonal effect is any variation in data due to calendar related effects which occur systematically at specific seasonal frequencies every year.

See: Time Series Data

Seasonally adjusted series

A seasonally adjusted series involves estimating and removing the cyclical and seasonal effects from the original data.

See: Time Series Data

Skewness (skewed distribution)

Skewness is the tendency for the values to be more frequent around the high or low ends of the x-axis.

See: Measures of Shape

Standard deviation

The standard deviation measures the spread of the data around the mean.

See: Measures of Spread

Standard error (SE)

The standard error (SE) is a measure of the variation between any estimated population value that is based on a sample rather than true value for the population.

See: Measures of Error

Statistical literacy

Statistical literacy refers to the knowledge and skills that enable data users and producers to understand, evaluate and communicate statistical data and information.

Statistical standard

A statistical standard is a set of rules used to standardise the way data are collected and statistics are produced.

See: What are Standards?

Statistic

A statistic is a value that has been produced from a data collection, such as a summary measure, an estimate or projection. Statistical information is data that has been organised to serve a useful purpose.

See: What are Statistics?

Stock series

A stock series is a measure of certain attributes at a point in time and can be thought of as “stock takes”.

See: Time Series Data

Survey

A survey involves collecting information from every unit in the population (a census), or from a subset of units (a sample) from the population.

See: Data Sources

T

Time series

A time series is a collection of observations of well-defined data items obtained through repeated measurements over time.

See: Time Series Data

Trend series

A trend series is a seasonally adjusted series that has been further adjusted to remove irregular effects and 'smooth' out the series to show the overall 'trend' of the data over time.

See: Time Series Data

V

Variable (data item)

A variable is any characteristic, number, or quantity that can be measured or counted.

See: What are Variables?

Variance

The variance measures the spread of the data around the mean.

See: Measures of Spread

Return to Statistical Language Homepage

This page last updated 22 October 2013